Санкт-Петербургский государственный

университет

Амяга Владислав Васильевич

Выпускная квалификационная работа

МОДЕЛИ SMART DATA В БИЗНЕС-АНАЛИЗЕ

Направление: 38.04.05 «Бизнес-информатика: Информационная бизнес-аналитика»

Основная образовательная программа магистратуры: «Информационная бизнес-аналитика»

Научный руководитель:
доцент, кандидат физико-математических наук
Гадасина Людмила Викторовна

_____/Подпись/

Рецензент:

_____/Подпись/

Санкт-Петербург
2018

Оглавление

1 Введение

1.1 Анализ предметной области

Современное развитие информационных технологий, относительная доступность информации, а также простота генерации новых данных рядовыми пользователями сформировали основные тенденции усиливающегося роста глобального объёма данных. Так, по результатам отчёта IDC, в 2011 году всеобщий объём произведённых и скопированных данных составил 1.8 ZB, увеличившись примерно в 9 раз от уровня 2006 года [1]. По прогнозам аналитиков глобальный объём данных будет удваиваться по крайней мере каждые два года в ближайшем будущем.

Традиционные информационные системы в условиях постоянно усиливающегося потока поступающих данных оказались не способны эффективно решать задачи, которые перед ними ставит бизнес. Эта проблема вызвана недостатками традиционных ИС и СУБД, среди которых можно выделить отсутствие гибкости, небольшие возможности к масштабированию и неэффективность при работе с различными типами данных.

С другой стороны, ограничения накладывают и сами данные. Помимо большого объёма, современные данные также довольно сильно различаются по формату своего представления. Так, например, данные могут быть получены абсолютно из любых источников, включая интернет и устройства различного назначения, и иметь совершенно различный формат – от сенсорных данных, до видео- и аудиофайлов.

В связи с этим предпринимаются постоянные попытки разрешить возникающие противоречия, что привело к появлению и развитию концепции «больших данных» (Big Data), а также специальных методов и подходов для работы с такими данными. В настоящее время «большие данные» главным образом ассоциируется с огромными наборами данных. По сравнению с традиционными данными, Big Data включает в себя значительный объём неструктурированных данных и в большей степени нуждается в обработке в реальном времени.

Одно из определений, раскрывающих понятие «больших данных», дало глобальное консалтинговое агентство McKinsey & Company. С точки зрения данного агентства, «большие данные» - это наборы данных, размер которых не позволяет приобретать, обрабатывать и хранить их с помощью традиционного программного обеспечения.

Следует отметить, что во многом это также зависит от отрасли экономики, которую мы берём для рассмотрения. Так, например, в зависимости от уровня используемых программных продуктов объём «больших данных» может варьироваться от нескольких терабайт (TB) до нескольких петабайт (PB) [2].

Понятие «большие данные» впервые было определено в 2001 году в исследовании Doug Laney, аналитика META [3]. В своей работе данный исследователь, сформулировал возможности и вызовы, которые открываются при возрастании объёма данных в соответствии с 3V моделью. Рост данных определялся 3 аспектами – объёмом (volume), скоростью (velocity) и разнообразием (variety).

Однако огромный объём данных сам по себе не несёт никакой пользы для компании, поэтому в 2011 году компания IDC расширила данную модель до модели 4 V, в которую дополнительно включила понятие ценности (value) [1].

Таким образом, понятие «больших данных» может быть определено четырьмя основными характеристиками:

- Большим объёмом данных (data volume)
- Требованиями к высокой скорости обработки данных (data velocity)
- Разновидностью поступающих данных (data variety)
- Ценностью, содержащейся в данных (data value)

В концепции Big Data объём и скорость определяют количественные аспекты информации, а разновидность и ценность – качественные.

Следует отметить, что некоторые аналитические компании расширяют предложенную модель до 7 V, дополнительно добавляя понятия изменчивости (variability), достоверности (veracity) и визуализации (visualization) [45]. Однако в данной научной работе мы будем пользоваться моделью 4 V.

Повышенное внимание к данной теме со стороны компаний объясняется их стремлением получить выгоду, возникающую при анализе Big Data [46]. Инвестиции и усилия, направленные на анализ «больших данных», полностью оправдывают себя, так как при объединении внутренних и внешних источников данных открываются огромные возможности по поиску ценности и совершению новых открытий. Так, например, анализ Big Data позволяет принимать нестандартные управленческие решения, гибче реагировать на реакции покупателей и создавать по-настоящему ценные продукты.

Выделим основные преимущества, которые открываются перед компаниями, использующими Big Data [30]:

- возможность привлечения и удержания клиента с самыми низкими затратами для компании
- управление взаимодействием с клиентом на оптимальном уровне рентабельности

- возможность относиться к каждому клиенту как к личности с уникальными вкусами, предпочтениями и ценностями
- возможность предсказывать поведение клиентов и основные тенденции рынка
- возможность исследовать скрытые отношения и зависимости
- значительное снижение расходов на рекламу
- снижение уровня риска

Вкладывая значительные средства и усилия в развитие Big Data, многие компании до сих пор с трудом извлекают из «больших данных» какую-либо ощутимую выгоду. В первую очередь это связано с технологическим и управленческим аспектами работы с «большими данными».

Изложенный выше аспект позволяет говорить о своевременности научных исследований по данной тематике, обосновывает ответ на вопрос, почему представленные проблемы должны быть изучены именно сейчас.

Так, на основании проведённого нами анализа среди главных проблем при работе с Big Data были выделены следующие:

1) Нерешённость вопроса, связанного с управлением данными – управление данными является компетенцией ИТ специалистов или менеджеров?
2) Потеря актуальности данных в связи с устареванием – «большие данные» довольно быстро становятся неактуальными и вследствие этого теряют какую-либо ценность, поэтому важным аспектом в деятельности компаний является своевременный анализ входящих потоков информации, в том числе анализ в режиме реального времени
3) Бизнес-процессы не адаптированы под работу с «большими данными» – при использовании «больших данных» возникает необходимость изменения большинства бизнес-процессов компании и переход на модель data-driven (управление данными)
4) Проблема взаимодействия ИТ и бизнеса – вследствие технологической сложности Big Data и отсутствия решения проблемы управления данными, анализ «больших данных» выполняют ИТ специалисты. Однако вопросы, на которые необходимо найти ответ в данных, способны сформулировать только бизнес-пользователи
5) Вопросы компетенций рядовых пользователей – бизнес-пользователи различного уровня не способны воспринимать «большие данные» без предварительной обработки; особенности представления «больших данных», а также существующие технологии их обработки требуют специальных технических знаний для

эффективной работы с Big Data. Поэтому проводить работу по анализу данных способны только квалифицированные специалисты по данным (data scientists). В результате этого процесс анализа Big Data становится привилегией исключительно ИТ специалистов, что значительно сужает потенциальные возможности для аналитики

6) Конфиденциальность «больших данных» – отсутствие законодательства, регламентирующего использование «больших данных». Наличие нерешённых вопросов, связанных с защитой и использованием персональных данных

Следует отметить, что помимо перечисленных частных проблем в сфере Big Data, необходимо также уделить особое внимание современным тенденциям развития и проблемным областям исследований в области в целом. Среди главных современных вызовов можно выделить следующие [6]:

1) Представление данных – большинство поступающих данных имеют определённый уровень гетерогенности (неоднородности) по типу, структуре, семантике, организации и доступности. Цель представления данных заключается в корректном отображении данных для последующего компьютерного анализа и обеспечении корректной интерпретации пользователем полученных результатов. Неправильное представление данных может значительно снизить ценность исходной информации и полностью свести на нет эффективность проводимого анализа
2) Сокращение избыточности и сжатие данных – современные ИТ технологии значительно упрощают создание новых данных. Развитие интернета вещей (IoT) и технологий облачных вычислений привело к возникновению проблем, связанных со сбором, интеграцией, управлением и обработкой широкого спектра данных из огромного числа распределённых источников. Вследствие этого, эффективное сжатие данных в предположении, что сокращение избыточности не окажет негативного влияния на ценность обрабатываемой информации, открывает широкие возможности для существенного сокращения затрат предприятия на хранение и обработку данных
3) Управление жизненным циклом данных – ценность, скрытая в Big Data, зависит от актуальности анализируемых данных, поэтому в настоящее время существует потребность в разработке принципов аналитической обработки данных для определения какие данные следует хранить, а какие данные должны быть отброшены

4) Аналитические механизмы – аналитические алгоритмы Big Data должны быть способны обрабатывать большое количество разнородных данных в течении ограниченного периода времени (в том числе и в режиме реального времени). Нереляционные БД (NoSQL) показали свои уникальные преимущества в обработке неструктурированных данных и в настоящее время являются основным инструментом для анализа «больших данных». Однако несмотря на это существуют определённые проблемы при использовании нереляционных БД, что выражается в потребности поиска компромиссного решения между традиционными СУБД и нереляционными БД. Необходимо также проведение более масштабных исследований по направлению использования оперативных БД (in-memory database) и методов, основанных на приближённом анализе (approximate analysis)
5) Информационная безопасность данных - многие компании, работающие с Big Data, в настоящее время не могут эффективно обрабатывать и анализировать огромные наборы данных из-за ограниченности собственных мощностей. В своём анализе большинство ИТ компаний должны полагаться на профессионалов или специальные сервисы для анализа «больших данных», что повышает потенциальные риски информационной безопасности. Поэтому анализ и обработка Big Data могут быть переданы третьей стороне только в том случае, если в целях обеспечения безопасности компании были приняты надлежащие превентивные меры по защите конфиденциальности данных
6) Расширение и масштабирование – аналитические алгоритмы «больших данных» должны быть способны обрабатывать увеличивающиеся и всё более и более усложняющиеся наборы данных
7) Кооперация - анализ Big Data это междисциплинарная область исследований, которая вынуждает экспертов различных сфер знаний сотрудничать друг с другом в целях увеличения потенциала использования «больших данных». Архитектура Big Data должна обеспечивать учёным и инженерам из различных областей науки доступ к широкому спектру информации, позволяя полностью использовать имеющийся опыт для достижения поставленных целей
8) Энергетический менеджмент – дальнейшее усовершенствование дата-центров (data center), рост объёма хранимых данных и аналитических потребностей в обработке, хранении и передачи «больших данных» неизбежно вызовет рост потребления электроэнергии. Таким образом, в целях дальнейшего расширения и масштабирования возможностей для анализа данных должны быть внедрены

действенные механизмы энергетического менеджмента и разработаны руководства для осуществления контроля потребления электроэнергии на уровне всей системы

Исходя из анализа предметной области исследований нами была выдвинута **гипотеза** о том, что в настоящее время в практической деятельности компаний существуют нерешённые проблемы в сфере анализа данных, получившие в академической литературе недостаточное освещение.

Таким образом, **цель** данной научной работы заключается в разработке модели бизнес-анализа на основе концепции Smart Data, обеспечивающей более эффективное использование данных в компании.

В свою очередь для достижения цели выпускной квалификационной работы, а также принятия или опровержения выдвинутой гипотезы необходимо решить следующие **задачи**:

- Обосновать актуальность, теоретическую значимость и прикладную ценность выбранного направления исследований
- Провести анализ современных научных работ
- Сформулировать основные нерешённые проблемы теории и практики, существующие в академической литературе
- Определить собственное место научных исследований
- Предложить вариант решения выявленных противоречий в виде разработанного решения (модели), удовлетворяющего потребности бизнеса

Объектом исследования настоящей работы является процесс бизнес-анализа в компании.

Предметом – модели, методы и инструменты в бизнес-анализе.

Методы проведения исследования – анализ, обобщение, сравнение.

1.2 Актуальность проведения исследований

Нерешённые проблемы практики, а также тенденции и перспективы развития области «больших данных» определяют необходимость и важность проведения дальнейших исследований по данной тематике. Возможности, которые открываются при работе с «большими данными», не могут быть использованы в полной мере по причине неразвитости бизнес-процессов, связанных с анализом и управлением Big Data. Отсутствие опыта и отработанных методик анализа и управления «большими данными» препятствует их эффективному использованию, что определяет необходимость разработки рекомендаций и готовых решений для разрешения существующих противоречий.

В настоящее время многие компании всерьёз задумались об изменении концепции и методов сбора данных, стремясь сделать этап отбора более осознанным и разумным. Одним из подходов, реализующих представленную идею, является концепция «интеллектуальных данных» (Smart Data).

Концепция Smart Data представляет собой подход по работе с данными, состоящий из двух аспектов:

1. Smart Data - данные, специально отобранные для решения конкретных задач и содержащие реальную бизнес-ценность для компании
2. Модель Smart Data - модель гибкой аналитики для получения и накопления бизнес-ценности (создание, наполнение и использование базы знаний компании)

Таким образом, Smart Data – это интерпретируемые человеком и приносящие бизнес-ценность существенные и актуальные данные с высокой степенью полезности и смыслового соответствия контексту.

Модель Smart Data – это форма представления и реализации концепции Smart Data, выраженная в виде процесса гибкой аналитики с обратной связью.

На рисунке 1.1 представлен обычный процесс работы с данными в компании (as is).

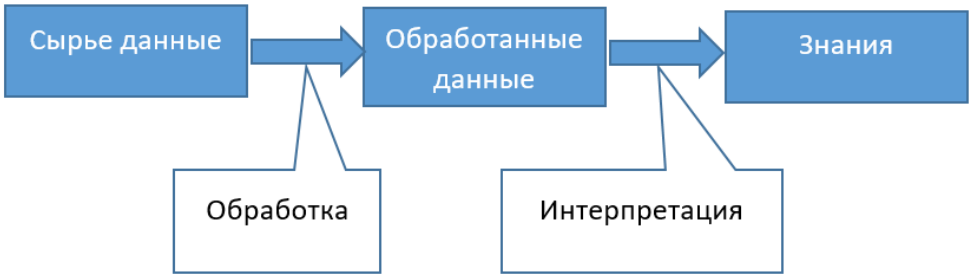

Рисунок 1.1. Процесс работы с данными as is

Основное отличие концепции Smart Data от обычного процесса работы с данными заключается в преимущественном отборе полезных данных, исходя из предыдущего опыта аналитики в компании. Таким образом, прошлый опыт анализа оказывает постоянное влияние на последующие методы и подходы работы с данными.

Процесс выстраивания аналитики в соответствии с концепцией Smart Data (as to be), представлен на рисунке 1.2.

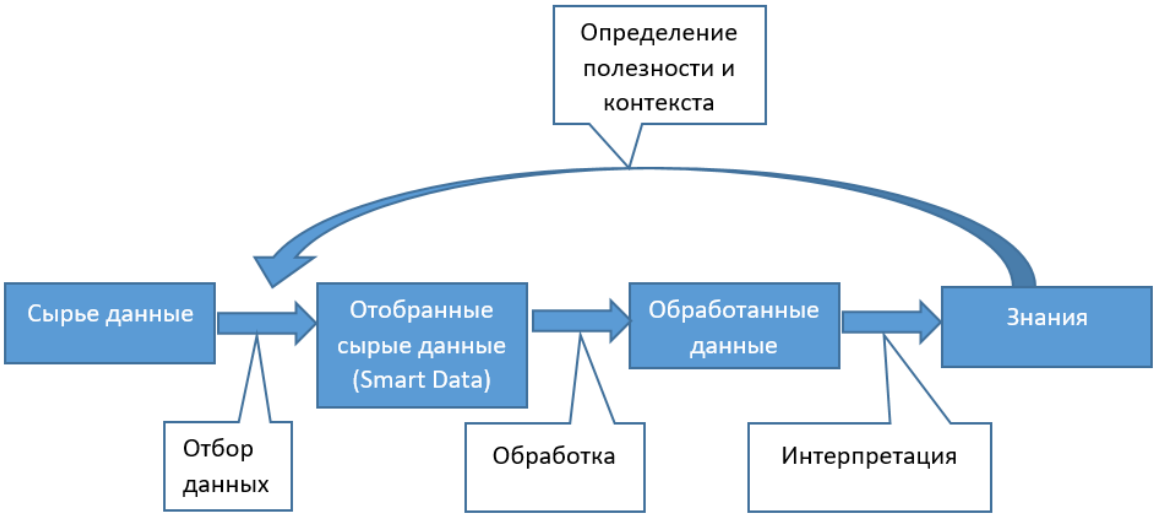

Рисунок 1.2. Модель Smart Data, представленная в виде процесса гибкой аналитики (as to be)

Так, например, разница между длинным списком чисел, относящихся к недельным продажам, по сравнению с выявлением пиков и впадин в течение долгого времени с помощью визуализации является частным случаем Smart Data. Сбор большого количества данных приносит мало пользы без дополнительного интеллектуального слоя.

Примерами знаний, извлекаемых из Smart Data, являются скрытые зависимости, тренды, выбросы и шаблоны, полученные в процессе интерпретации человеком результатов анализа данных. Реализация концепции Smart Data и использование интеллектуальных алгоритмов позволяет превращать бессмысленные цифры в легко интерпретируемую для человека визуальную информацию.

Среди основных свойств Smart Data можно выделить следующие:

1) Релевантные
2) Существенные
3) Актуальные
4) Практически полезные
5) Доступные для интерпретации

Информация из внешнего мира поступает из различных источников в структурированном и неструктурированном виде. В реальной жизни большинство данных содержат в себе информационный шум и не несёт никакой пользы для компании, поэтому зачастую машинные алгоритмы способны дать эффективные рекомендации по продукту, используя небольшие наборы данных. Сбор и хранение избыточного объёма данных значительно увеличивает затраты компании без соответствующего увеличения бизнес-ценности.

Вследствие этого при обработке Big Data возникает необходимость задавать вопросы, устанавливать цели сбора и обработки данных. В своих поисках компании могут выбирать два типа работы с данными:

6) Идти от сферы деятельности компании и конкретной практической задачи (тогда задача сбора и обработки данных существенно ограничивается, однако повышается точность предсказания и прогнозирования)
7) Идти от самих данных (так, компания применяет все возможные варианты обработки данных и в случае успеха способна находить ранее неизвестные зависимости и получать при этом значительные конкурентные преимущества)

Однако бесспорным фактом остаётся то, что для дальнейшего развития Big Data требуется скорость и интеллект. С каждой секундой человечество генерирует огромное количество данных, которые должны обрабатываться так же быстро. Следует отметить, что один только факт наличия большого количества данных недостаточен. Ключевые вопросы заключаются в том, являются ли они единообразными и регулярными, могут ли быть легко

извлечены и проанализированы, присутствует ли в них значительное количество вариаций, встроены ли данные в общую массу другой нерелевантной информации.

Это налагает определённые требования, предполагающие, что обработка и интерпретация Big Data не должна быть случайной. Сбор и использование данных имеет смысл только тогда, когда это используется для оптимизации бизнес-процессов, автоматизации и решения реальных проблем компании. Главная идея заключается не в бездумном сборе огромных массивов всевозможной информации, а в сборе каждого набора данных в рамках определённого контекста. Данные должны быть понятны и интерпретированы в рамках определённой ситуации. [40]

Методы концепции Smart Data ориентированы прежде всего на качественные аспекты данных, включающие в себя достоверность и ценность. Особое внимание также уделяется качеству и полезности анализируемой информации, результаты анализа которой в дальнейшем используются при принятии управленческих решений.

Вследствие этого концепция Smart Data подразумевает отбрасывание информационного шума и выделение наиболее ценной информации из всей совокупности данных. В дальнейшем в рамках определённой проблемы предполагается, что обработанная информация будет использоваться компаниями для решения конкретных поставленных задач.

Глобальный рост количества информации открывает широкие возможности извлечения ценности для бизнеса и более глубокого анализа скрытых зависимостей. Особую важность данной проблеме также придаёт и тот факт, что по мнению некоторых аналитических изданий, данные являются новой нефтью в цифровой экономике [41]. Грамотная работа с «большими данными» предполагает наличие эффективных навыков по организации, управлению и анализу поступающей информации.

В настоящее время существуют следующие подходы к решению проблем работы с данными в компаниях:

1. создание должности CDO-Chief Data Officer
2. применение интеллектуальных методов анализа данных
3. внедрение систем управления данными

Эффективная реализация концепции Smart Data предполагает, что компания должна иметь структуру, позволяющую проводить первичный отбор данных. Современные тенденции вынуждают компании более тщательно относиться к выбору стратегий управления данными, что проявляется в постоянном реагировании на изменяющиеся условия рынка. Исходя из возникающих потребностей, можно говорить о том, что любое изменение должно отвечать требованиям бизнеса и способствовать реализации целей компании. Поэтому закономерным

ответом на современные вызовы является учреждение в компаниях должности директора по данным (CDO-Chief Data Officer).

Следует отметить, что для того, чтобы отвечать потребностям бизнеса, а также реализовывать успешные бизнес-практики предполагается, что данную должность должен занимать человек не из ИТ подразделения, а из высшего звена управления компании. Следовательно, в компаниях всё чаще встаёт вопрос не о том, как собрать данные, а о том, зачем это нужно и для каких целей это использовать.

Ответственность CDO заключается в следующих аспектах деятельности:

- Управление корпоративными данными
- Развитие информационно-технологической архитектуры
- Внедрение инноваций в область управления данными

В современных условиях всё чаще компании используют огромный потенциал цифрового бизнеса, что проявляется в увеличении потребности качественного управления корпоративной информацией, а также в возрастании роли аналитики в операционной деятельности [42]. Поэтому введение должности CDO является закономерным результатом развития компаний и направлено на реализацию множества возможностей, возникающих в результате сбора информации в масштабах бизнеса.

С бурным ростом информационного потока важнейшей задачей, которую призван решать CDO, становится поиск и использование только необходимой информации, способной добавить ценность бизнесу, увеличить производительность предприятия или способствовать более качественному риск-менеджменту. Только специалист такого уровня способен в полной мере разобраться в потребностях компании с точки зрения использования данных и провести работу по переходу от концепции Big Data к Smart Data.

Однако многие CDO отмечают, что существует высокий уровень сопротивления изменениям со стороны ИТ-отделов, а также постоянная борьба за контроль над информационными ресурсами и их управлением. Поэтому при внедрении данной должности необходимо также учитывать и этот аспект.

По оценкам аналитических изданий темп прироста новой информации с каждым годом будет только увеличиваться, и огромное количество поступающих данных возможно будет обработать только с помощью новых методов и концепций, о которых раньше никто и не задумывался. Эффективное использование концепции Smart Data отлично подходит для решения данной задачи.

Все вышеперечисленные положения подтверждают актуальность и практическую значимость исследований по данной тематике.

Таким образом, для достижения цели и решения задач настоящей выпускной квалификационной работы необходимо найти ответы на следующие проблемные вопросы:

1) Как следует устранить противоречие между возможностями, которые открываются при работе с «большими данными», и ограничениями, накладываемыми практической деятельностью компании?
2) Как повысить эффективность работы компании при анализе «больших данных»?

Как уже было сказано выше, представленные проблемные вопросы являются актуальными и существует реальная потребность бизнеса в поиске рекомендаций для решения данных проблем.

2 Теоретическая часть

Для анализа современных направлений научных исследований по тематике «больших данных» и определения области Smart Data был проведён литературный обзор современных научных работ. В процессе проведения литературного обзора была подробно проанализирована предметная область исследований, классифицированы существующие направления научных работ, выявлены нерешённые проблемы теории и практики, а также предложены рекомендации по проведению дальнейших исследований.

Теоретическая часть настоящей выпускной квалификационной работы содержит методологический материал по теме исследования, литературный обзор современных научных работ, а также интеллектуальный анализ текста отобранных исследований.

2.1 Обзор литературы и анализ работ по близким к теме ВКР исследованиям

В литературном обзоре современных научных исследований были проанализированы наиболее значимые работы, посвящённые области «больших данных», а также затронуты исследования, касающиеся других смежных направлений - Smart Data, интеллектуальный анализ данных, машинное обучение, интернет вещей (IoT) и облачные технологии.

Цель данного обзора заключалась в определении современных направлений научных исследований, поиске проблемных областей, а также уточнении понятия Smart Data, складывающегося в современной научной литературе.

Методология проведения литературного обзора была основана на рекомендациях, предложенных в работах Boell и Cecez-Kecmanovic [4], а также Yichuan Wang [5].

Так, данный литературный обзор состоял из четырёх основных частей:

- Поиск и приобретение
- Преобразование и классификация
- Критическая оценка и аргументирование
- Разработка рекомендаций для дальнейших исследований

На этапе поиска и приобретения подходящих научных работ были использованы инструменты базы данных Scopus, а также следующий набор ключевых слов:

1) Big Data
2) Smart Data
3) Data Mining
4) Business Intelligence

Так, были проанализированы работы, соответствующие заданным ключевым словам в заголовке, а также временному периоду с 2012 по 2018 год. Отметим также, что в русскоязычных источниках зачастую отсутствуют аналоги терминов, встречающихся в зарубежной научной литературе, поэтому для проведения анализа современных направлений исследований «больших данных» были использовали англоязычные источники.

Критерием для дальнейшего отбора научных работ послужила их публикация в авторитетных научных изданиях, а также представление на научных конференциях по исследуемой тематике.

Таким образом, основными критериями отбора научных работ являлись следующие:

5) Соответствие работы отобранным ключевым словам
6) Соответствие работы предметной области исследований
7) Публикация работы в авторитетных научных журналах и сборниках конференций
8) Наибольший индекс цитирования по данной тематике

В результате были отобраны топ-10 научных работ в каждой тематике, ранжированных по индексу цитирования. Данная процедура повторялась для каждого набора ключевых слов. Так, в процессе поиска по ключевым словам отбор на соответствие заявленным критериям прошло 40 работ. Завершающим этапом поиска было удаление дубликатов, в процессе которого количество исследуемых работ сократилось до 24.

Результаты первичного отбора научных работ представлены в таблице 2.1

Таблица 2.1

Ключевые слова	Количество исследований, удовлетворяющих критериям	Количество исследований без дубликатов
Big Data	10	10
Smart Data	10	5
Data Mining	10	5
Business Intelligence	10	4
Всего	40	24

На этапе преобразования и классификации отобранные работы были распределены по различным тематикам в соответствии с тезисами, представленными в текстах работ. Отобранные работы с соответствующими идеями, выводами и вкладом в новые знания были классифицированы по различным областям, в которых также были определены ключевые направления исследований.

Далее, на этапе критической оценки и аргументирования был проведён анализ содержимого статей, уточнены современные направления научных работ, а также выделены проблемные места в текущих исследованиях. В настоящей выпускной квалификационной работе были выделены современные проблемы, с которыми сталкиваются специалисты в

области анализа «больших данных». На основании анализа Min Chen, Shiwen Mao и Yunhao Liu [6], а также ряда других исследователей были классифицированы современные тенденции развития сферы «больших данных».

В результате проведённого анализа было выделено пять основных направлений исследований:

- Решение текущих технологических проблем
- Развитие и совершенствование новых технологических направлений
- Управленческий аспект «больших данных»
- Правовой аспект «больших данных»
- Междисциплинарный аспект «больших данных»

Решение текущих технологических проблем

При сборе данных из большого числа разнородных источников возникает проблема их хранения и интеграции. Отсутствие единых правил представления данных значительно усложняет анализ и ухудшает качество его результатов. Это связано с тем, что разнородность данных может запутывать исследователей, вследствие чего существующие зависимости и значимые факторы не будут найдены. Поэтому для повышения эффективности анализа необходимо решить проблемы сбора, хранения и визуализации данных различных форматов и представлений.

Также должны быть решены вопросы, связанные с масштабируемостью, доступностью и целостностью данных. Одним из вариантов решения представленных проблем является дальнейшее развитие и совершенствования средств и инструментов бизнес-аналитики, а также переход к нереляционным БД.

Актуальными проблемами в рамках данного направления также является сокращение избыточности и сжатие данных, вследствие того, что большое количество информации, поступающей с датчиков, не несёт в себе никакой ценности и значительно усложняет анализ. Нерешёнными также остаются и вопросы преобразования данных и контроля их качества.

Развитие и совершенствование новых технологических направлений

Одним из наиболее активно развивающихся направлений в области «больших данных» является технологии облачных вычислений. Облачные вычисления – это технология распределённой обработки данных для выполнения сложных и больших по масштабу вычислений. Компьютерные ресурсы и вычислительные мощности предоставляются пользователям как интернет-сервис, что позволяет отказаться от необходимости покупки, размещения и поддержания дорогостоящего вычислительного оборудования.

Облачные технологии становятся одним из ключевых факторов роста и развития производственных компаний. Применение данной технологии способствует трансформации традиционной бизнес-модели предприятия, обеспечивает согласование продуктовых инноваций, а также позволяет создавать интеллектуальные производственные сети, способствующие эффективному взаимодействию сотрудников и разработке новейших бизнес-стратегий.

Также в настоящее время широкое распространение получило направление «интернета вещей» (IoT), в том числе концепция «умных» городов, нацеленная на использование самых передовых коммуникационных технологий в целях поддержания услуг с добавленной ценностью для администрации города и граждан.

Отметим также, что вместе с ростом потребности в вычислительных мощностях, вызванной необходимостью обработки огромного количества разнородных данных, в крупных информационно-технологических компаниях начала активно развиваться концепция «дата-центров», что является одним из наиболее перспективных направлений исследований.

Управленческий аспект «больших данных»

Развитие облачных технологий, интернета вещей, концепции «умных» городов и других направлений информационных технологий стимулируют компании к изменениям своей структуры и внутренней информационной среды. Так, например, многие предприятия внедряют новые подходы к управлению данными, что выражается в развитии концепции «data-driven» и ориентации на применение информационных технологий в каждом бизнес-процессе компании.

В то же время актуальность поступающей информации является важнейшей составляющей бизнес-ценности, поэтому для поддержания высокого уровня конкурентоспособности компании необходимо качественное управление жизненным циклом данных.

Вследствие этого возникают задачи сохранения и поддержания долговечности ретроспективных данных компании. Частично эти проблемы можно решить с помощью создания «дата-центров», однако вместе с этим возникают вопросы, связанные с интеграцией и управлением данными. Также в больших масштабах «дата-центров» важнейшей проблемой становится разработка методов проведения грамотного энергетического менеджмента.

Правовой аспект «больших данных»

Научные исследования в рамках данного направления затрагивают вопросы неприкосновенности частной жизни и правовые аспекты регулирования в сфере информационного права и персональных данных.

Помимо частного аспекта защиты персональных данных, прорабатываются и пути для охраны корпоративной информации, связанные с их защитой и сохранением конфиденциальности.

Междисциплинарный аспект «больших данных»

Научные работы по тематике «больших данных» затрагивают исследовательскую и практическую деятельность в различных областях науки – биологии, физике, экономике, астрономии, государственном управлении, социологии, медицине и др. Практически невозможно представить современное научное исследование, которое бы не сталкивалось с обработкой больших массивов данных. Вследствие этого возникают закономерные вопросы взаимодействия и кооперации со специалистами из различных предметных областей и научных направлений.

2.2 Интеллектуальный анализ текста

В рамках данного литературного обзора для более точного описания текущих направлений исследований, а также выявления существующих противоречий был проведён интеллектуальный анализ текста. Для анализа текста использовались встроенные возможности языка R, а также ряд пакетов и библиотек, специализирующихся на анализе текстовых данных. Опустим технический процесс проведения исследования и перейдём непосредственно к результатам.

В результате анализа 24 работ, отобранных на этапе поиска и приобретения, были выделены слова и ключевые словосочетания, наиболее часто встречающиеся в исследуемой литературе. Результаты представлены на рисунке 2.1.

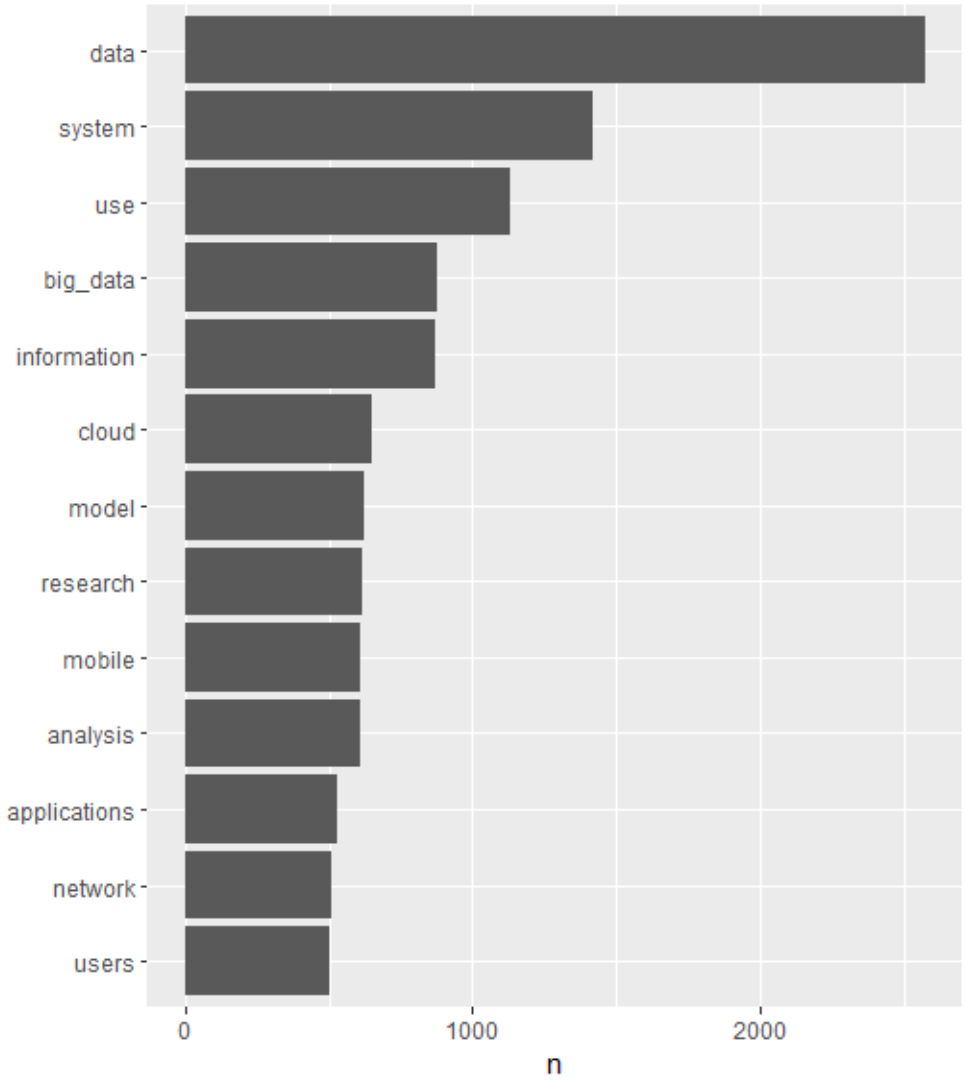

Рисунок 2.1. Рейтинг наиболее часто встречающихся слов

При анализе линейчатой диаграммы, представленной на рисунке 2.1, нельзя сделать однозначные выводы, так как большинство отобранных слов встречаются практически во всех научных работах по тематике «больших данных».

Вследствие этого, для более точного анализа, нас интересовала частота встречаемости ключевых терминов, которые были использованы для отбора научных работ на этапе поиска и приобретения. На рисунке 2.2 представлены результаты поиска данных ключевых словосочетаний в текстах отобранных научных работ.

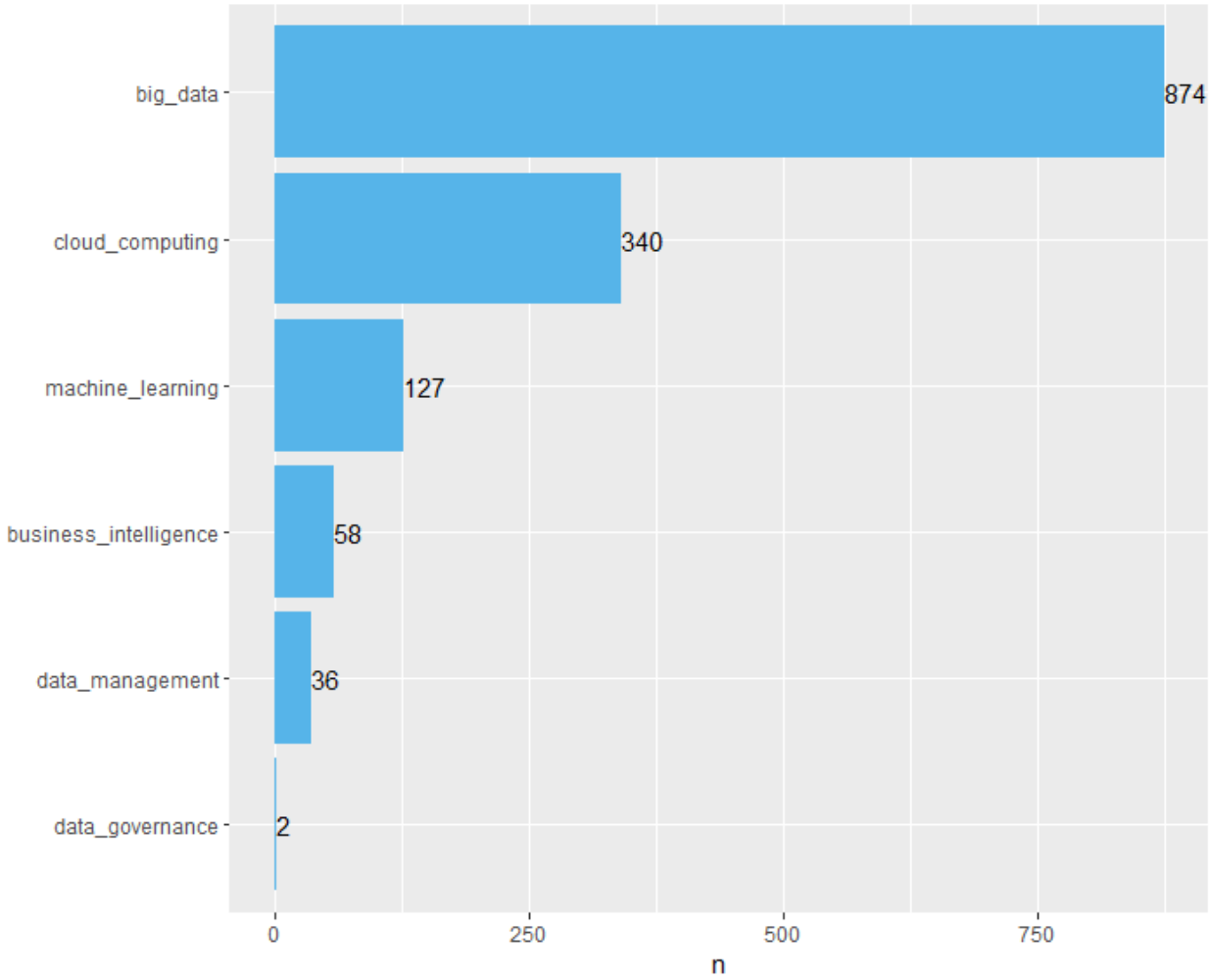

Рисунок 2.2. Частота ключевых словосочетаний

Анализ диаграммы, представленной на рисунке 2.2, позволяет выявить ключевые области исследований, которым уделено наибольшее внимание в современной научной литературе, а также отразить проблемные или недостаточно освещённые области исследований.

Как мы можем заметить, ключевыми технологиями при работе с «большими данными» являются облачные вычисления (cloud computing) и машинное обучение (machine learning).

В то же время наблюдается значительный недостаток исследований в области управления данными. Так, на рисунке 2.3 словосочетания «data management» и «data governance» находятся на последнем месте в представленном рейтинге, что говорит о недостаточной освещённости данной тематики в передовой научной литературе.

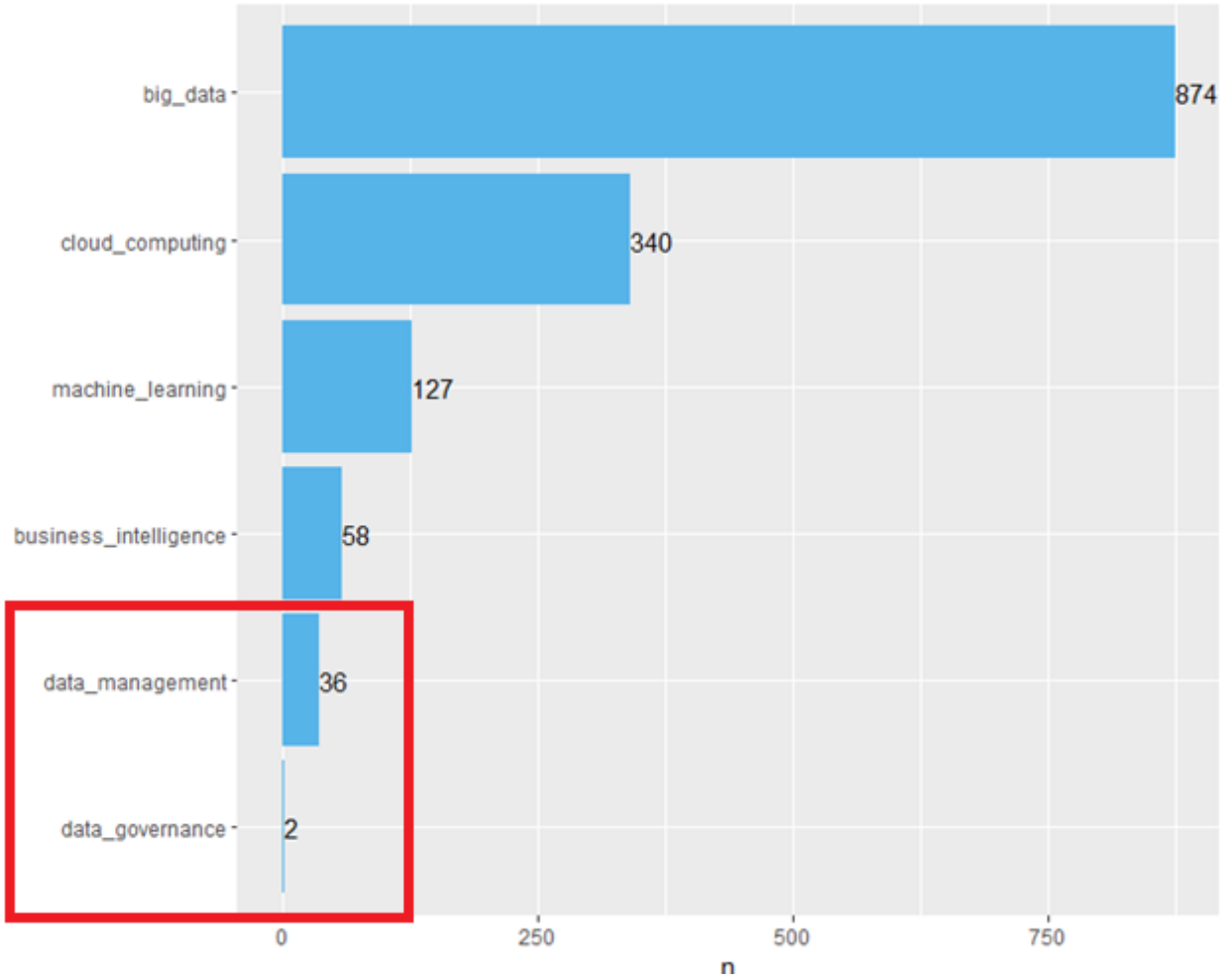

Рисунок 2.3. Частота ключевых словосочетаний

Данный факт в совокупности с выводами, полученными из введения настоящей выпускной квалификационной работы, позволяет сделать вывод о том, что в научной литературе уделено недостаточное внимание проблемам управления «большими данными». Однако, как уже было сказано выше, проблема качественного управления данными является одной из ключевых в практической деятельности компаний, что говорит о высокой потребности со стороны бизнеса в проведении исследований по данной теме.

Далее, для определения характера проблем, существующих в области «больших данных», был проведён анализ настроений. На рисунке 2.4 отражены результаты данного анализа.

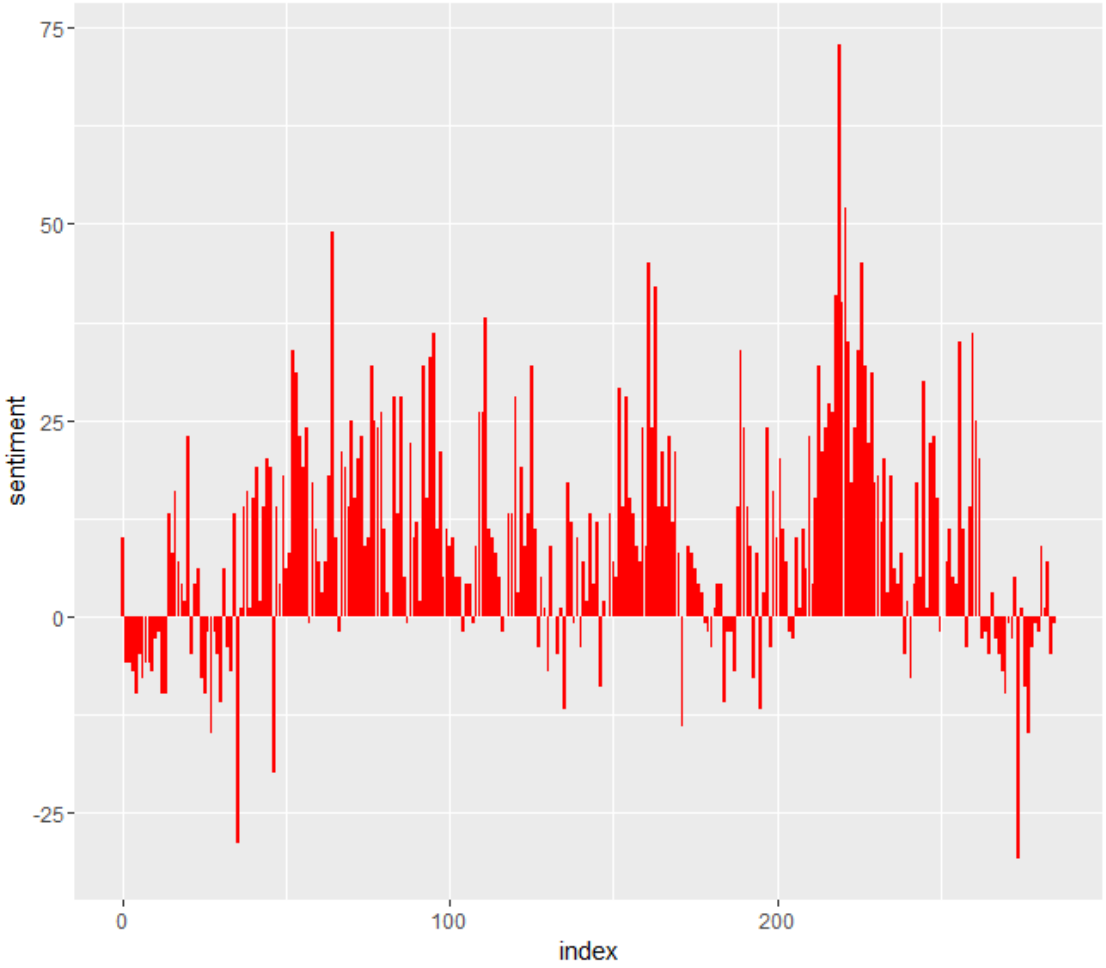

Рисунок 2.4. Анализ настроений научных работ

В ходе проведения данного анализа в текстах анализируемых документов были выделены слова, содержащие настроение и на их основе проведена бинарная классификация на позитивные и негативные слова. В целом можно сделать вывод о том, что подавляющее большинство исследований по теме «больших данных» носит позитивный характер.

Этот факт также же иллюстрирует рисунок 2.5. На данной диаграмме выделен рейтинг слов, которые встречаются в текстах научных статей наибольшее количество раз. Так, красным были выделены негативные слова, синим-позитивные.

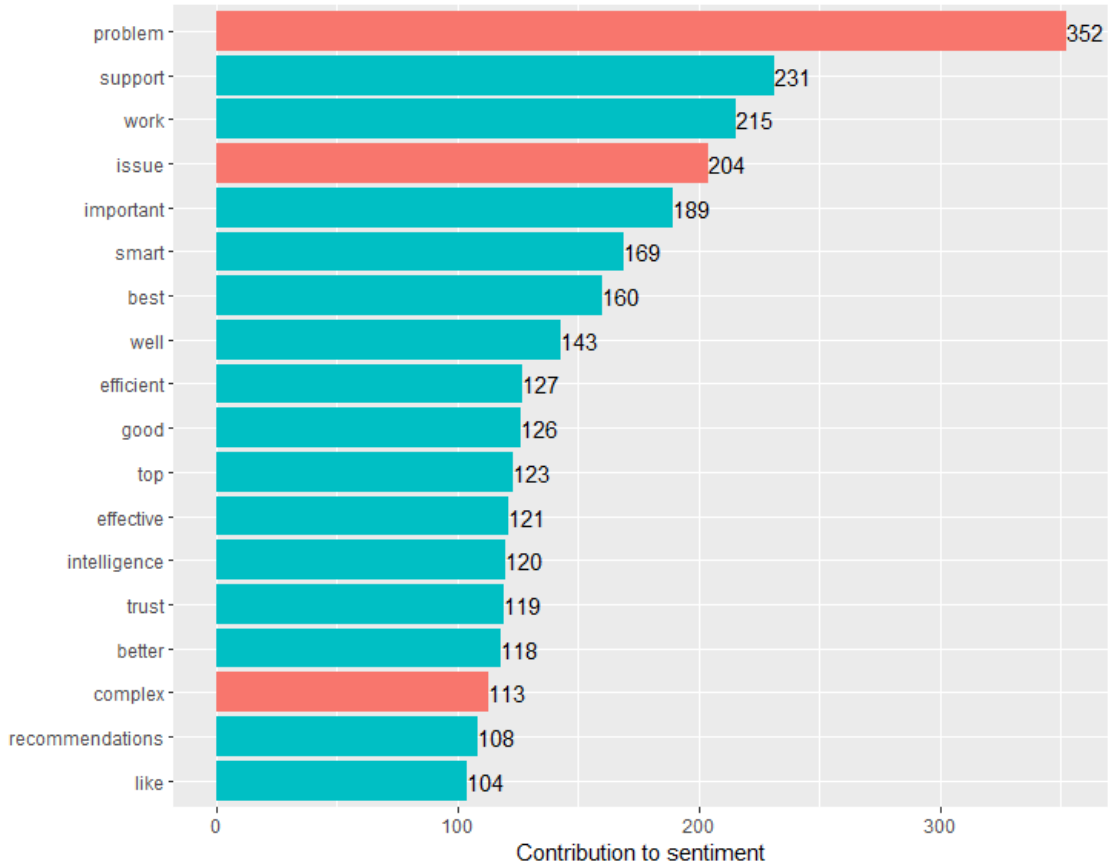

Рисунок 2.5. Позитивные и негативные слова

Для большей наглядности количественных различий выведем облако слов. Облако слов представлено на рисунке 2.6.

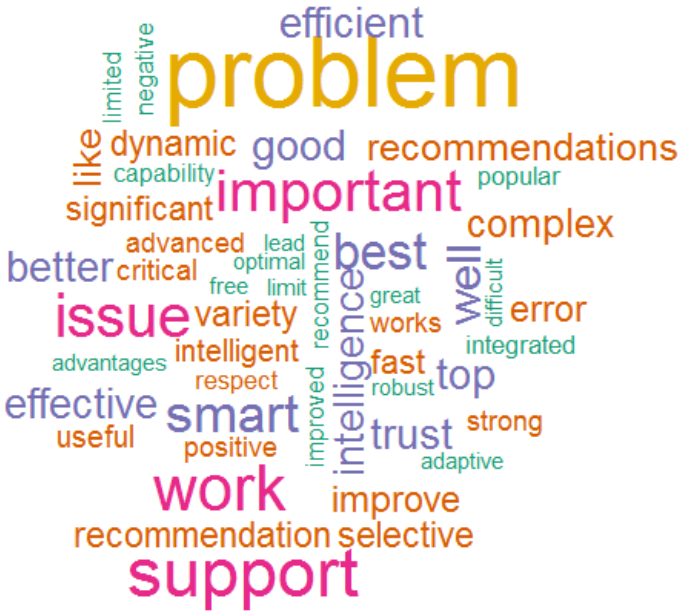

Рисунок 2.6. Облако слов

Далее, распределим слова на негативные и позитивные. Результат классификации представлен на рисунке 2.7.

Рисунок 2.7. Облако позитивных и негативных слов

Таким образом, можно заметить значительно выделяющиеся по сравнению с остальными слова «problem» (проблема) и «issue» (вопрос), что говорит о перспективности и актуальности рассматриваемых направлений научных исследований, а также о наличии в них большого числа проблем и вопросов, требующих дальнейшей проработки и поиска возможных решений.

2.3 Выводы

В результате проведения литературного анализа современных исследований были определены области, недостаточно освещенные в отобранных научных работах. Среди данных проблем были выделены следующие:

1. Практически отсутствуют исследования, связанные с применением ИТ в государственных органах управления
2. Исследования носят преимущественно прикладной характер, теоретические аспекты исследований недостаточно широко освещены
3. Недостаточно освещена тема трансформации профиля экономики и структуры компании вследствие использования технологий обработки «больших данных»
4. Недостаточно освещена тема управления «большими данными» в компании
5. Отсутствие в академической литературе проработанных подходов для работы с данными в соответствии с концепцией Smart Data

Последний этап литературного обзора заключался в разработке рекомендаций и предложений для проведения дальнейших научных исследований. В целом рекомендации касаются вопросов и проблемных областей, рассмотренных выше. На наш взгляд, области исследований, которые получили недостаточную освещённость в проанализированных научных работах являются перспективными и также требуют внимания исследователей.

Особо отметим проблемы, связанные с управлением данными. В связи с тем, что новейшие ИТ технологии изменяют профиль современных компаний и значительно модернизируют их внутренние бизнес-процессы, возникает закономерная потребность в разработке качественной и эффективной методологии управления данными. Эта работа носит всеобщий характер, так как охватывает многие сферы деятельности компании, начиная от процесса управления и заканчивая непосредственно производством и операционными бизнес-процессами. По результатам проведённого анализа можно сделать вывод о том, что в настоящее время существует недостаточная проработанность вопроса управления «большими данными», несмотря на то, что данная тема является одной из важнейших в деятельности компании. Сформулированная проблема требует дальнейшей проработки и соответствующих исследований.

Как было выявлено раннее, технологический аспект не является основной проблемой в концепции «больших данных». Тем не менее специалисты отмечают, что существуют проблемы с интеграцией информационных систем различного рода. Следствиями данных проблем являются сложности при переносе накопленной информации из старых систем в новые, а также проблемы поддержания систем, основанных на принципах обработки

«больших данных» в режиме реального времени. В связи с этим весьма значительной является проблема управления данными и, связанный с ней тесным образом, вопрос поддержания качества данных.

По мнению экспертов, политика управления данными должна основываться на анонимности, прозрачности, сохранении конфиденциальности и справедливой стоимости использования. Так, например, большинство людей согласны предоставлять свои персональные данные в разумных пределах, если компании компенсируют это с помощью скидок, бонусов или приятных дополнений. Клиенты начинают чувствовать себя некомфортно, если они вынуждены предоставлять свои персональные данные просто так, без какой-либо компенсации.

В некоторых аналитических отчётах делается попытка очертить области применения «больших данных», а также обозначить основные вызовы и проблемы, с которым ещё предстоит столкнуться исследователям и практикам данной предметной области. Так, например, всё больше специалисты начинают говорить не о размере данных, а о том, что компании делают с ними, обращая особое внимание на способность аналитиков правильно задавать вопросы и выдвигать гипотезы к данным. С точки зрения большинства экспертов проблемы состоят не в технической плоскости (какими методами и технологиями обрабатывать данные), а в концептуальной (различного рода организационные, культурные и идеологические проблемы).

Большинство исследователей также отмечает вызовы в сфере применения «больших данных», с которыми только предстоит столкнуться в будущем. Компаниям, которые пошли по пути использования «больших данных» придётся приложить изрядные усилия для того, чтобы найти решение будущих проблем.

Отметим, что современные научные исследования не дают чёткого определения концепции Smart Data. Данная ситуация является следствием того, что понятие «больших данных» до сих пор не устоялось и появилось сравнительно недавно, поэтому часть исследователей относит Smart Data к отдельному направлению развития «больших данных» и не выделяет для него отдельной терминологии. Другая же часть исследователей, напротив, несколько смело заявляет о закате Big Data и появлении нового направления Smart Data, так как «большие данные» в настоящее время не отвечают современным вызовам и требованиям. Однако и в первом и во втором случае чётких определений никто не даёт.

Современные тенденции обозначают и будущие источники данных – различного рода датчики и сенсоры (в компьютерах, телевизорах, холодильниках и т.д.), умные пространства, носимые устройства, биометрическое и медицинское оборудование. Из вышеперечисленного

можно сделать вывод, что количество данных о различных аспектах человеческой жизни будет с каждым годом только расти [30].

Таким образом, в процессе проведения литературного обзора и последующего анализа текста были определены современные направления исследований в области «больших данных», выявлены недостатки и проблемные места текущих исследований, а также предложены рекомендации для проведения последующих научных работ. Предложенная методология проведения литературного обзора позволила выделить ключевые направления развития научных исследований по теме «больших данных», предоставила возможность определения основных неразрешённых проблем, а также послужила основой для обоснования актуальности современных исследований по тематике Smart Data.

Так, результатом литературного обзора стало выявление проблем недостаточной освещённости темы управления данными, а также недостаточной теоретической проработанности подходов для работы с данными в соответствии с концепцией Smart Data. В рамках данного обзора была обоснована актуальность проведения исследований по тематике Smart Data, а также предложены возможные направления дальнейших научных работ.

3 Предпроектный анализ компании

В настоящей главе выпускной квалификационной работы была рассмотрена деятельность конкретной компании, проведён анализ её информационно-технологической инфраструктуры, исследован процесс аналитики и выявлены текущие проблемы в сфере анализа данных, тесно пересекающиеся с основными противоречиями, выделенными в теоретической части данной работы.

3.1 Описание сферы деятельности компании

Для успешного выявления текущих проблем компании и информационных потребностей бизнес-пользователей необходимо проведение комплексного анализа особенностей предприятия. Данный анализ состоит в подробном описании бизнес-процессов, ИТ-инфраструктуры, а также сферы деятельности компании.

Основные виды деятельности компании, в рамках которой была проведена данная научная работа, заключаются в разведке и разработке месторождений нефти и газа, нефтепереработке, а также производстве и сбыте нефтепродуктов. Структурно компания является вертикально-интегрированной и подразделяется на блоки и дочерние общества. Каждый из блоков компании ответственен за свое поле деятельности.

В рамках данной выпускной квалификационной работы была рассмотрена деятельность блока логистики, переработки и сбыта, в частности процесса нефтепереработки.

3.2 ИТ-инфраструктура компании

Современные тенденции развития ИТ технологий, которые были подробно описаны в теоретической части, а также возрастание информационных потребностей бизнеса, привели к формированию и развитию концепции «Озера данных».

«Озеро данных» - способ хранения данных в системе или в хранилище в естественном формате, в целях облегчения обработки данных в различных структурных формах [47]. Идея «Озера данных» состоит в создании единого хранилища всех данных на предприятии, начиная от необработанных (точные копии исходных системных данных) до преобразованных данных, используемых для задач отчетности, визуализации, аналитики и машинного обучения. «Озеро данных» включает в себя инструменты пакетной и потоковой обработки информации и может содержать в себе хранилище данных.

«Озеро данных» состоит из:

6. Структурированных данных (данные из реляционных баз данных)
7. Полуструктурированных данных (csv, журналы, xml, json)
8. Неструктурированных данных (электронные письма, документы, PDF-файлы, видео, аудио, изображения)

Среди ключевых факторов, стимулирующих развитие и внедрение данной концепции выделяются следующие:

1. Рост объемов корпоративных данных:
    a. Данные с датчиков (интернет вещей)
    b. Данные партнеров и клиентов
    c. Внешние источники данных и дата-сервисы
2. Появление новых источников неструктурированных данных:
    a. Электронная почта
    b. СМС и мессенджеры
    c. Данные веб-аналитики
    d. Отзывы клиентов в социальных сетях
    e. Геоданные
    f. Анализ видео и фото
3. Появление новых требований к бизнесу:
    a. Быстрая и недорогая обработка больших объемов данных (в том числе и неструктурированных)
    b. Доступность всех типов данных для разнообразных инструментов обработки
    c. Максимально быстрая проверка гипотез, возникающих по поводу данных
    d. Недорогое хранение большого объема данных (тенденция сохранить как можно больше данных в исходном, необработанном виде)
    e. Гибкость при интеграции новых источников данных
    f. Наличие открытых интерфейсов (API) для получения и предоставления данных (возможность обогащения и монетизации данных)

В свою очередь благодаря технологиям распределенного хранения и обработки концепция «Озера данных» успешно решает данные проблемы.

Основные преимущества данной концепции по сравнению с классическим хранилищем данных (data warehouse) заключаются в следующем:

1. «Озеро данных» в отличие от традиционных ХД хранит все данные предприятия. Данный подход становится возможным, поскольку аппаратные средства для функционирования «озёр данных» отличаются от оборудования, используемого для ХД. Использование распределенных хранилищ данных позволяет значительно снизить затраты на хранение данных, при этом в процессе работы никакие данные не удаляются.
2. «Озеро данных» поддерживает хранение всех типов данных.

   Данные хранятся как есть, трансформации происходит только в том случае, когда появляется необходимость в использовании данных. Такой подход известен как «Schema on Read» в отличие от «Sheme on Write», реализующейся в ХД.
3. «Озеро данных» поддерживает все типы пользователей.

   Пользователи, участвующие в анализе данных:
   a. Бизнес-пользователи - их деятельность полностью удовлетворяется ХД и включает в себя создание отчетов, проверку стандартных показателей
   b. Бизнес-аналитики - совершают больше экспериментов с данными, для них возможностей ХД уже недостаточно, поэтому они часто во время анализа обращаются к системам первоисточникам
   c. Ученые по данным (Data Scientists) - используют внешние источники данных, зачастую игнорируют ХД, так как работают с совершенно различными типами данных, не содержащихся в ХД

   Подход озера данных поддерживает всех этих пользователей. Ученые по данным могут использовать «большие данные», которые хранятся в «озере», обычные же пользователи могут использовать более структурированные данные, для удовлетворения своих информационных потребностей.
4. «Озеро данных» легко адаптируется к изменениям

   В отличие от ХД, «озеро данных» хранит данные в необработанном виде, и пользователи могут исследовать данные по-новому. Если в процессе анализа выяснится, что результаты исследований полезны, то к «озеру данных» всегда возможно применить более структурированную схему, автоматизацию и повторное использование для распространения результатов на более широкую аудиторию.
5. «Озеро данных» способствует более быстрому пониманию и выявлению зависимостей

Хранение данных в необработанном виде позволяет совершать более быстрые исследования данных и выявление зависимостей в отличии от традиционного подхода ХД.

Тем не менее, зачастую сложно осуществить моментальный перенос всех данных из существующего и заполненного ХД в «озеро данных». Одним из возможных решений данной проблемы может быть применение гибридного подхода, сочетающего в себе архитектуру ХД и «озера данных».

Так, в компании, в рамках которой была проведена данная работа, реализована гибридная архитектура информационных систем, БД и компонентов управления данными. Схематичное описание данной архитектуры представлено на рисунке 3.1.

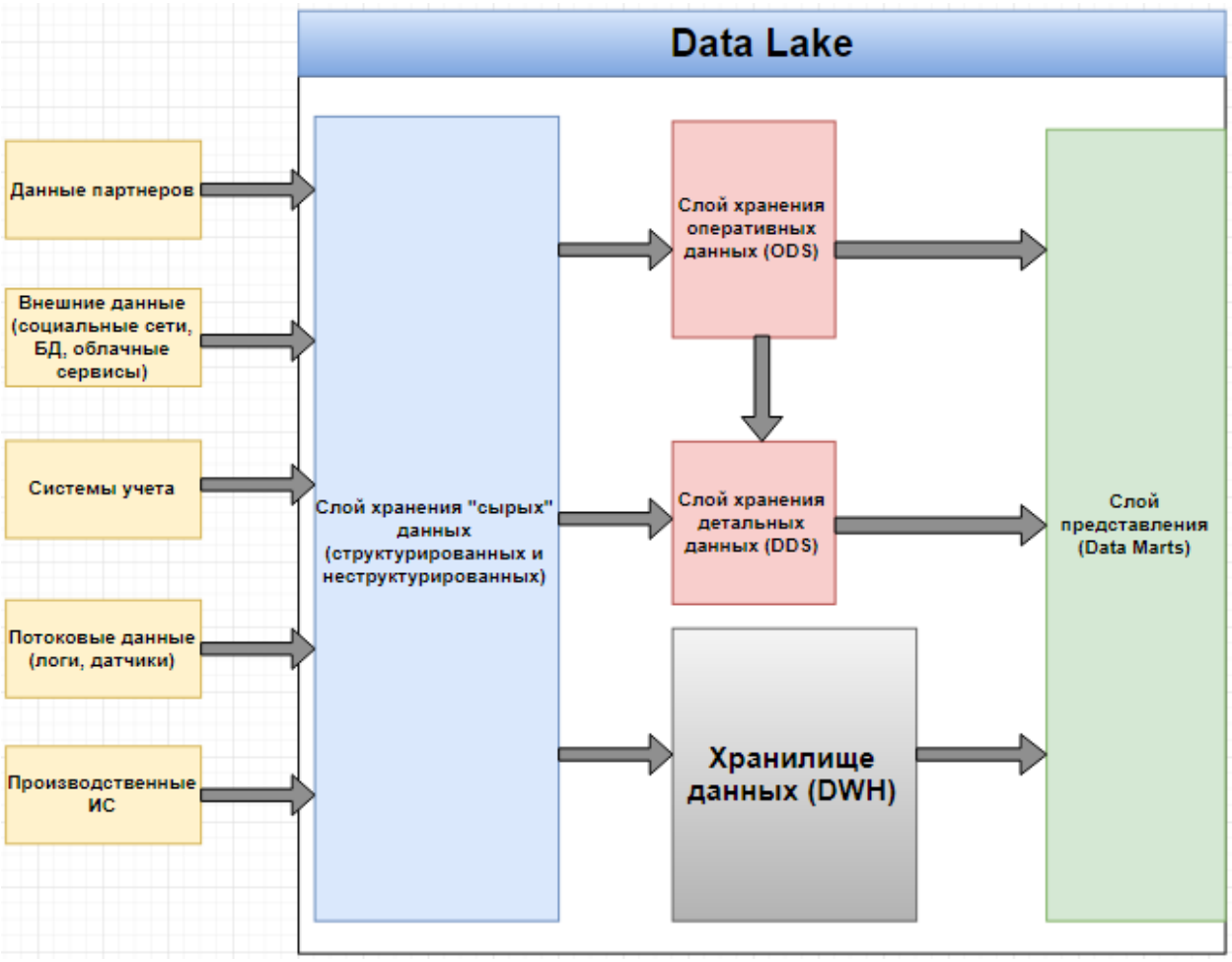

Рисунок 3.1. Гибридная схема ИТ-инфраструктуры

Тем не менее, реализация гибридного подхода несет в себе определенные проблемы и риски. Более подробно эти проблемы, а также рекомендации по их разрешению будут описаны в следующих главах данной выпускной работы.

3.3 Описание общего процесса аналитики в компании

Процесс аналитики в компании в общем виде состоит из следующих основных этапов:

1. Извлечение сырых данных
    a. Объединение нескольких источников данных
    b. Извлечение релевантных данных
2. Предобработка
    a. Трансформация данных - преобразование и консолидация данных в форму, пригодную для обработки и применения методов интеллектуального анализа данных
    b. Очистка данных - удаление зашумленности и несоответствий
3. Загрузка данных в ХД
4. Интеллектуальный анализ данных - применение интеллектуальных методов и алгоритмов для извлечения скрытых шаблонов, трендов и выбросов
5. Интерпретация
    a. Оценка шаблонов - определение значимой и ценной информации
    b. Представление данных - визуализация добытых знаний для пользователей
    c. Выводы по поводу данных
    d. Получение знаний

Таким образом, в общем виде процесс аналитики в компании представлен на рисунке 3.2.

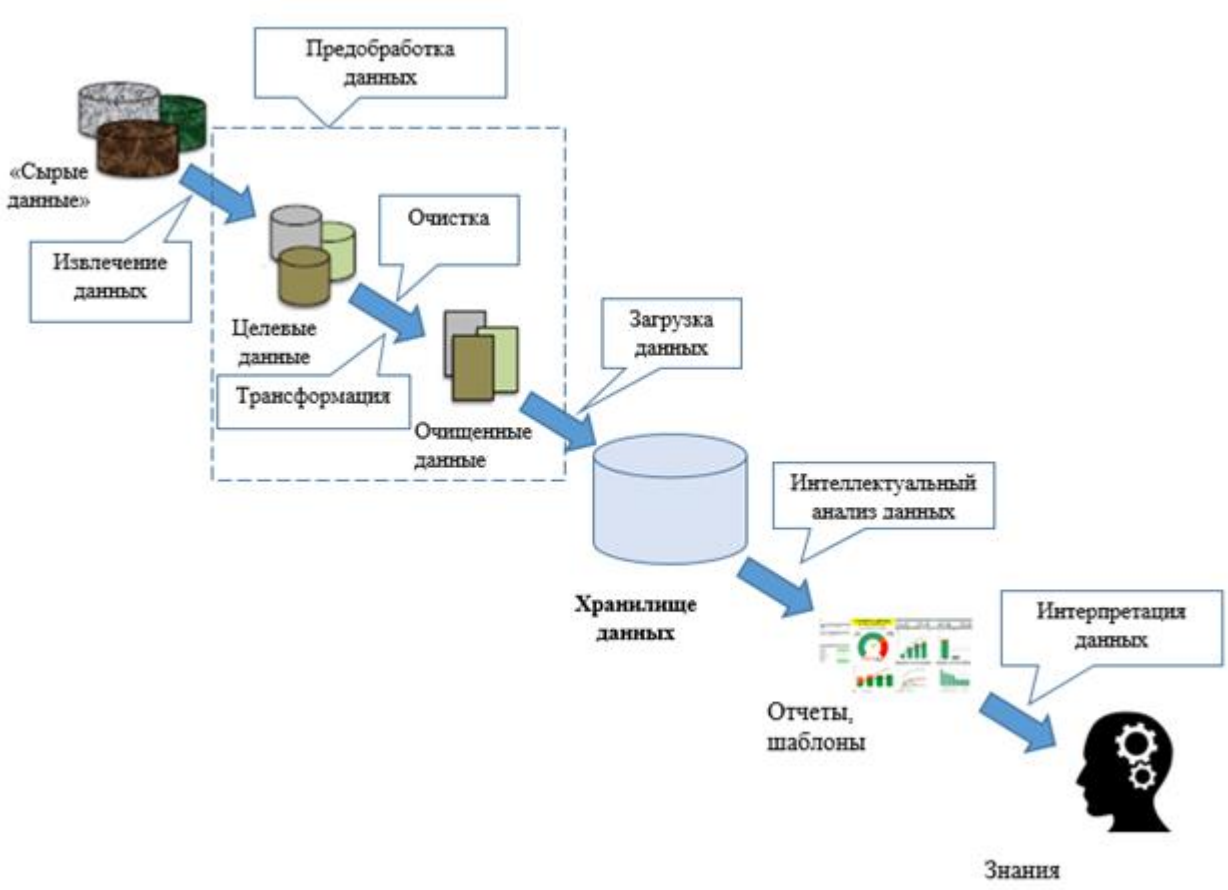

Рисунок 3.2. Процесс аналитики в компании

Заметим, что при реализации гибридного подхода концепции «озера данных» на этапе загрузки данных в ХД могут возникнуть проблемы при интеграции различных источников. Вследствие этого должна быть проведена работа по качественному управлению данными, их обогащению, созданию реестра потоков данных, а также полной, единой и непротиворечивой системы нормативно-справочной информации (далее НСИ). В противном случае на этапе интеллектуального анализа данных высока вероятность получения результатов низкого качества и, вследствие этого, принятие неверных решений и совершение управленческих ошибок.

Представленный процесс аналитики в компании заканчивается на этапе интерпретации данных, однако для повышения эффективности работы предприятия и сохранения лучших практик должен быть налажен постоянный процесс накопления полученных знаний. Таким образом, процесс накопления экспертизы не должен быть стихийным и случайным – для решения данной задачи необходимо выстроить четкий процесс формирования базы знаний компании и проработать дальнейший процесс управления полученными знаниями.

3.4 Описание проекта внедрения системы аналитической отчетности по цепочке добавленной стоимости

Для последующего анализа существующих проблем мы ограничимся областью деятельности предприятия, связанной с анализом эффективности нефтепереработки. Данная область деятельности компании состоит из комплексного анализа работы установок нефтеперерабатывающего завода, контроля качества поступающего сырья и вырабатываемой продукции.

В настоящий момент бизнес-пользователям, участвующим в процессе анализа эффективности нефтепереработки, на постоянной основе приходит оперативная информация в виде ежедневных отчетов:

1. отчет о работе установок (план, факт и отклонение материальных балансов)
2. сводка о выработке нефтепродуктов (план и факт по смешению, паспортизация товарной продукции)
3. сводка о качестве нефти и нефтяного сырья

Также у пользователей есть доступ к информационным системам, из которых они ежедневно выгружают необходимую им информацию по качеству потоков и технологическим режимам установок.

Процесс анализа цепочек добавленной стоимости в свою очередь построен следующим образом:

1. Бизнес-пользователь просматривает отклонения материальных балансов компонентов товарной продукции (до смешения)
2. Если было выявлено отклонение, то для выяснения причины, осуществляется мониторинг всей цепочки установок, участвующих в формировании интересующего компонента. На основании этого определяется системность отклонения по конкретной установке
3. Далее, просматривается качество потоков в рамках всей цепочки производства интересующего компонента товарной продукции и определяется соответствие исходного сырья нормам по качеству
4. После завершения анализа бизнес-пользователь делает выводы о причинах отклонений

Таким образом, причина отклонений может заключаться в плохом качестве исходного сырья или в отклонении работы конкретной установки.

Подводя итог вышесказанному, можно выделить следующие активности в рамках данного анализа:

1. Анализ качества поступающего сырья (нефть и нефтепродукты)
2. Анализ отклонений качества потоков и товарной продукции
3. Анализ работы установок нефтеперерабатывающего завода (показатели по фактическому выходу потоков, технологические показатели работы установок)
4. Анализ отклонений фактического смешения компонентов продукции
5. Анализ отклонений фактической выработки товарной продукции

Так как для работы бизнес-пользователи используют большое количество различных источников, процесс аналитики является трудоемким. Вследствие этого, руководством предприятия было принято решение о разработке приложения QlikView «Система аналитической отчетности по цепочке добавленной стоимости» для консолидации всех источников данных в единую систему отчетности и налаживания комплексного и последовательного процесса аналитики, значительно упрощающего ежедневную работу бизнес-пользователей.

В процессе разработки приложения QlikView будут использованы следующие данные:

- Качество потоков и нефтепродуктов (хранятся в системе LIMS)
- Информация по технологическим режимам установок (содержится в системе PI)
- План, факт по материальным балансам потоков (записываются в систему АСКУБ)
- Первоисточники по плановым показателям качества и материальным балансам (хранятся в PIMS)
- Сводки по качеству нефти и нефтяного сырья (на ежедневной основе заполняются в Excel)

Данные из представленных источников будут загружаться в хранилище данных и на следующем этапе использоваться в приложении QlikView.

На рисунке 3.3 отображено схематичное описание элементов ИТ-инфраструктуры, участвующих в представленном процессе аналитики.

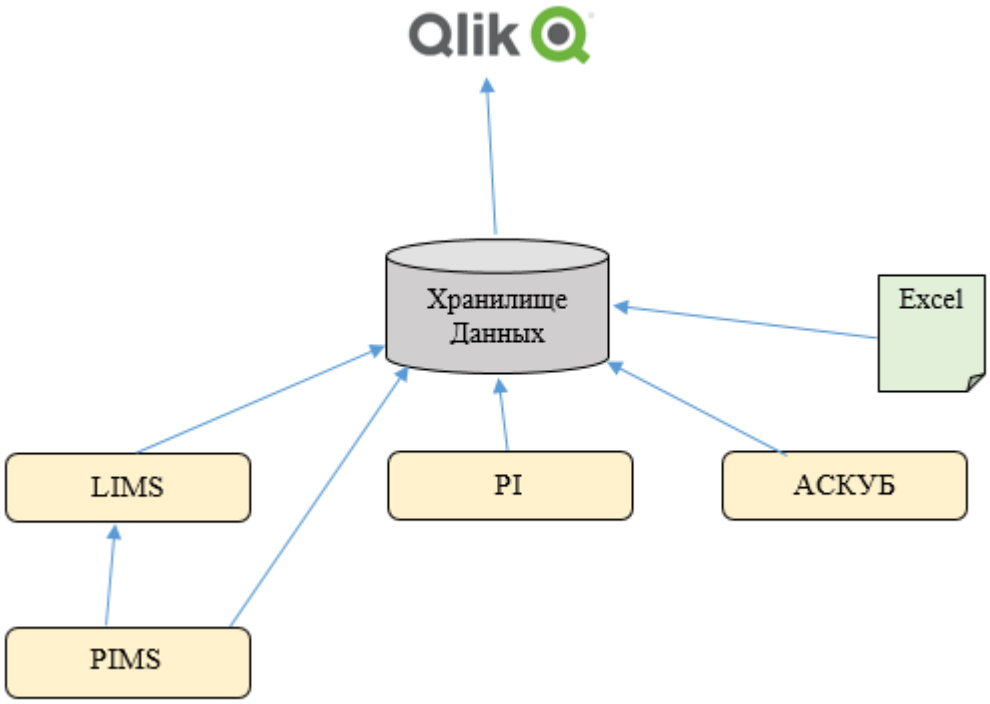

Рисунок 3.3. Информационные системы, данные из которых используются в процессе аналитики

Следует также отметить, что разработка целевого приложения ведётся в соответствии с гибкой методологией разработки Agile.

На рисунке 3.4 представлена вкладка «Нефть» одной из версий разрабатываемого целевого приложения по анализу цепочек добавленной стоимости.

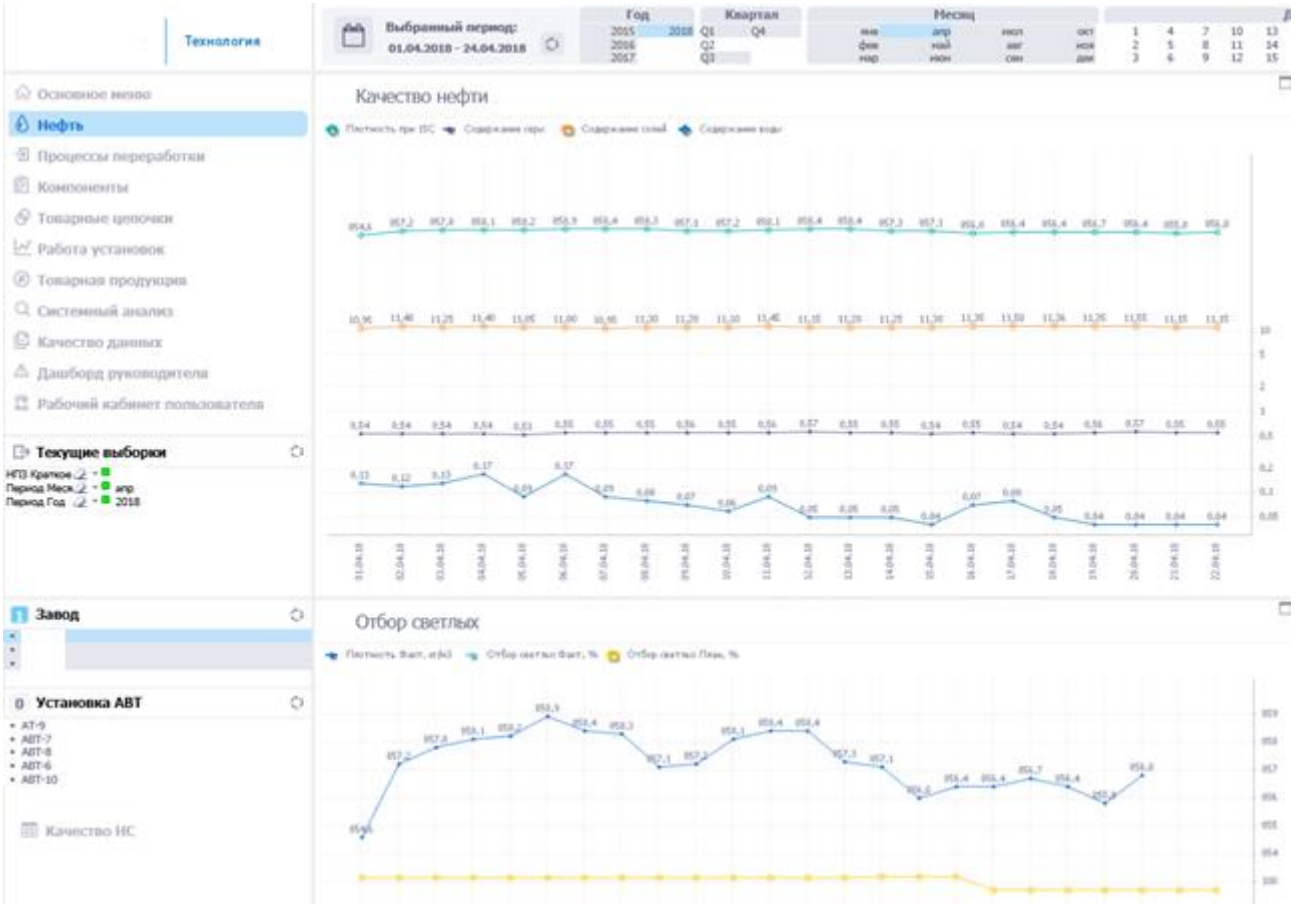

Рисунок 3.4. Вкладка «Нефть» целевого приложения

Процесс разработки целевого приложения производится по следующим ключевым этапам:

1. Планирование и анализ требований
2. Проектирование
3. Разработка и внедрение версии приложения
4. Использование пользователями приложения в повседневной работе
5. Сбор требований и анализ обратной связи от пользователей
6. Подготовка информации для следующего этапа разработки

Исходя из данного процесса разработки приложения, у руководителя проектной команды зачастую возникают вопросы об эффективности выпускаемых версий приложений. Для руководства важно получить ответы на следующие вопросы:

1. улучшила ли новая версия приложения работу пользователей?
2. были ли достигнуты поставленные цели текущего этапа разработки?

Таким образом, решение данной задачи и поиск ответов на поставленные вопросы, значительно повысит качество разрабатываемого продукта, улучшит уровень сервиса и окажет поддержку проектной команде разработчиков.

3.5 Выявление текущих проблем компании

Как уже было отмечено выше, при реализации концепции «озера данных» на этапе загрузки данных в ХД зачастую проявляются проблемы интеграции различных источников. В то же время в больших компаниях, активно использующих ИС, неизбежно возникают трудности, связанные с управлением данными.

Основные причины этих проблем заключаются в следующем:

3. Большое количество независимых и не связанных между собой ИС
4. Сложность поддержки и управления большого количества ИС
5. Проблема доступности и безопасности данных
6. Организационные проблемы и политики компании

Таким образом, при рассмотрении общего процесса аналитики в компании эта проблема возникает на этапе загрузки данных в ХД, а затем проявляется на этапе интеллектуального анализа данных и напрямую влияет на качество получаемых результатов.

Также, напомним, что процесс аналитики в компании не должен заканчиваться на этапе интерпретации полученной информации. Для повышения эффективности работы предприятия и сохранения лучших практик должен быть налажен постоянный процесс накопления полученных знаний.

Выявленные проблемы в процессе аналитики представлены на рисунке 3.5.

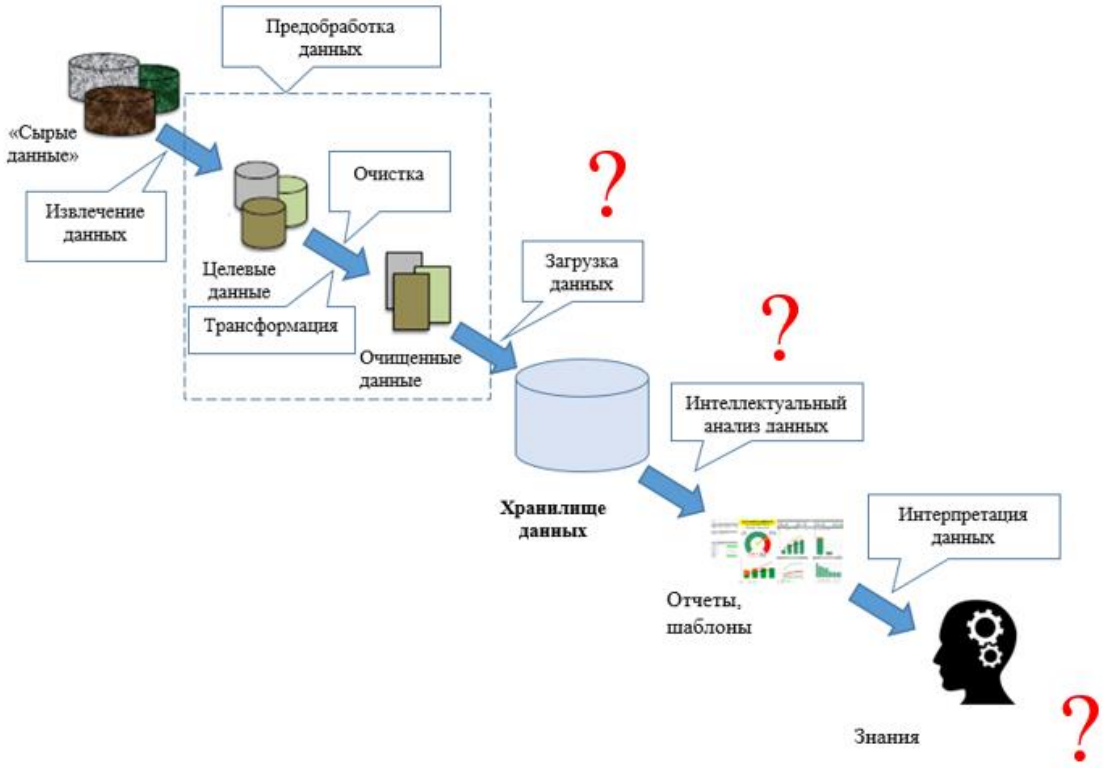

Рисунок 3.5. Проблемные места в процессе аналитики

Помимо прочего, у руководства существует реальная потребность в оценке эффективности разрабатываемого приложения QlikView «Система аналитической отчетности по цепочке добавленной стоимости». Отсутствие понимания удовлетворенности пользователей может привести к тому, что внедряемое приложение попросту не будет использоваться, что приведет к провалу проекта внедрения.

Косвенно оценить эффективность приложения для пользователей можно с помощью анализа лог-файлов, генерируемых самим приложением QlikView. В данных лог-файлах содержится информация об активности бизнес-пользователей в приложении, их действиях, наиболее часто выбираемых элементах и прочая полезная информация.

Проблема заключается в том, что анализировать данную потоковую информацию на постоянной основе способны лишь пользователи с высоким уровнем технических знаний. Как правило, данным анализом с помощью специальных методов занимаются ИТ-специалисты, в то время как правильные вопросы могут сформулировать только пользователи от бизнеса.

Вследствие этого, возникает реальная потребность разработки приложения, на постоянной основе представляющего данную информацию в удобном и воспринимаемом виде.

Таким образом, в процессе реализации проекта «Системы аналитической отчетности по цепочке добавленной стоимости» QlikView были выявлены следующие проблемы:

1. Наличие большого количества источников и отсутствие единой системы НСИ (в том числе и единых подходов по грамотному управлению данными)
2. Отсутствие налаженного процесса консолидации экспертизы и полученных знаний, возникающих в процессе ежедневного анализа
3. Реальная потребность бизнеса в анализе потоковых данных для понимания эффективности разрабатываемого приложения

4 Реализация концепции Smart Data на примере конкретной компании

В данной главе настоящей выпускной квалификационной работы была реализована концепция Smart Data на примере решения конкретных проблем компании, выявленных в предыдущей главе данной работы.

4.1 Разработка рекомендаций по единой системе НСИ

Отсутствие единой системы НСИ значительно усложняет процесс анализа данных, а иногда сводит на нет все затраты и усилия вложенные в процесс бизнес-аналитики.

Выделим основные проблемы, которые возникают при отсутствии единой системы НСИ:

- Неоднозначная интерпретация информации
- Значительные трудозатраты по поддержанию множества отдельных справочников сопоставлений
- Сложность, а иногда и невозможность интеграции данных из различных систем

Процесс поддержания справочников при отсутствии единой НСИ представлен на рисунке 4.1.

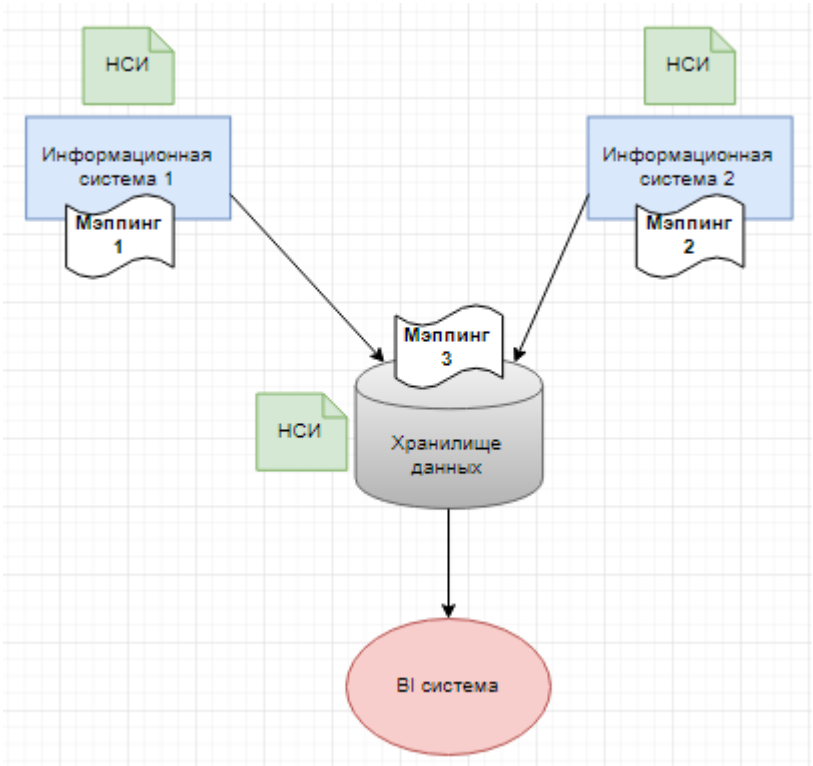

Рисунок 4.1. Схема сопоставления данных из различных систем при отсутствии единой НСИ

Как мы можем заметить, при увеличении количества систем растет и количество справочников, а также таблиц сопоставлений. Возникают проблемы по поддержанию и актуализации разнородных справочников.

При реализации единого подхода к созданию и управлению НСИ возможно избежать большинство из описанных проблем. Так, на рисунке 4.2 представлена схема сопоставления данных из различных систем при наличии единой системы НСИ.

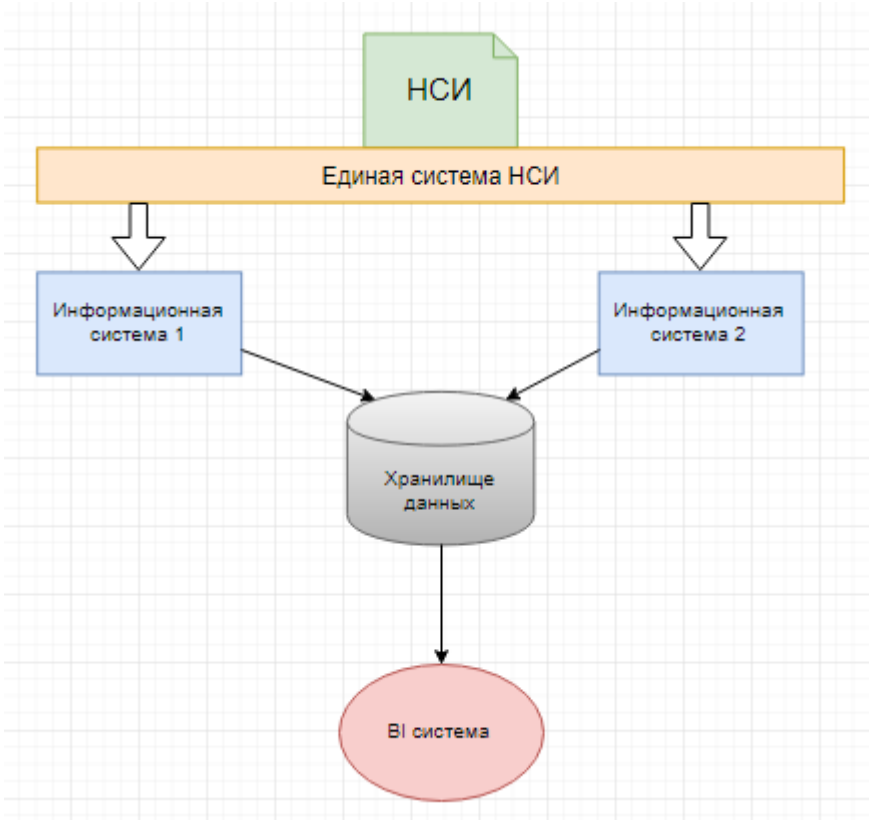

Рисунок 4.2. Схема сопоставления данных из различных систем при наличии единой НСИ

Как мы можем заметить, при такой схеме управления данными удается значительно сократить количество справочников и таблиц сопоставлений, а также избежать затрат на их постоянную актуализацию.

В то же время единая НСИ несет значительные выгоды для бизнеса, среди которых можно выделить следующие:

1. Однозначное и непротиворечивое представление информации
2. Обеспечение стабильности и непрерывности работы ИС
3. Повышение надежности интеграции данных между системами
4. Улучшение процесса управления и контроля

5. Единая система показателей эффективности компании
6. Синхронизация справочных данных со смежными процессами и документами
7. Создание единого механизма поддержки НСИ, единого места хранения, снижение трудозатрат на поддержание справочников сопоставлений
8. Увеличение скорости работы с данными
9. Значительное снижение трудозатрат при интеграции данных из различных систем
10. Возможность составления консолидированной отчетности
11. Возможность объединять данные, увеличивая их ценность
12. Повышение качества исходных данных
13. Снижение риска потери данных
14. Сокращение времени на внесение изменений в объекты НСИ

Так, в компании, в которой была проведена данная научная работа, разработана трехуровневая система НСИ, представленная на рисунке 4.3.

Рисунок 4.3. Архитектура НСИ

Данная архитектура позволяет делегировать управление данными на нижние уровни при сохранении общей целостности.

Состав данной архитектуры:

1. Корпоративный уровень – корпоративные системы управления нормативно-справочной информацией
2. Уровень блока – нормативно-справочная информация отдельного блока компании
3. Уровень подразделения и дочерних обществ

При реализации данной архитектуры основные рекомендации по процессу управления НСИ состоят в последовательном выполнении следующих этапов:

1. Паспортизация показателя – описание порядка формирования показателя, выделения всех справочников, которые используются для расчетов и отображения информации
2. Сбор и систематизация информации о справочниках – регистрация справочников (описание, назначение), определение владельцев, уточнение методики и регламента ведения и поддержания справочников
3. Создание эталонных справочников и справочников сопоставлений - интеграция с целевыми системами, реализация возможности дублирования исходных справочников и обеспечение передачи эталонов

Данный процесс представлен на рисунке 4.4.

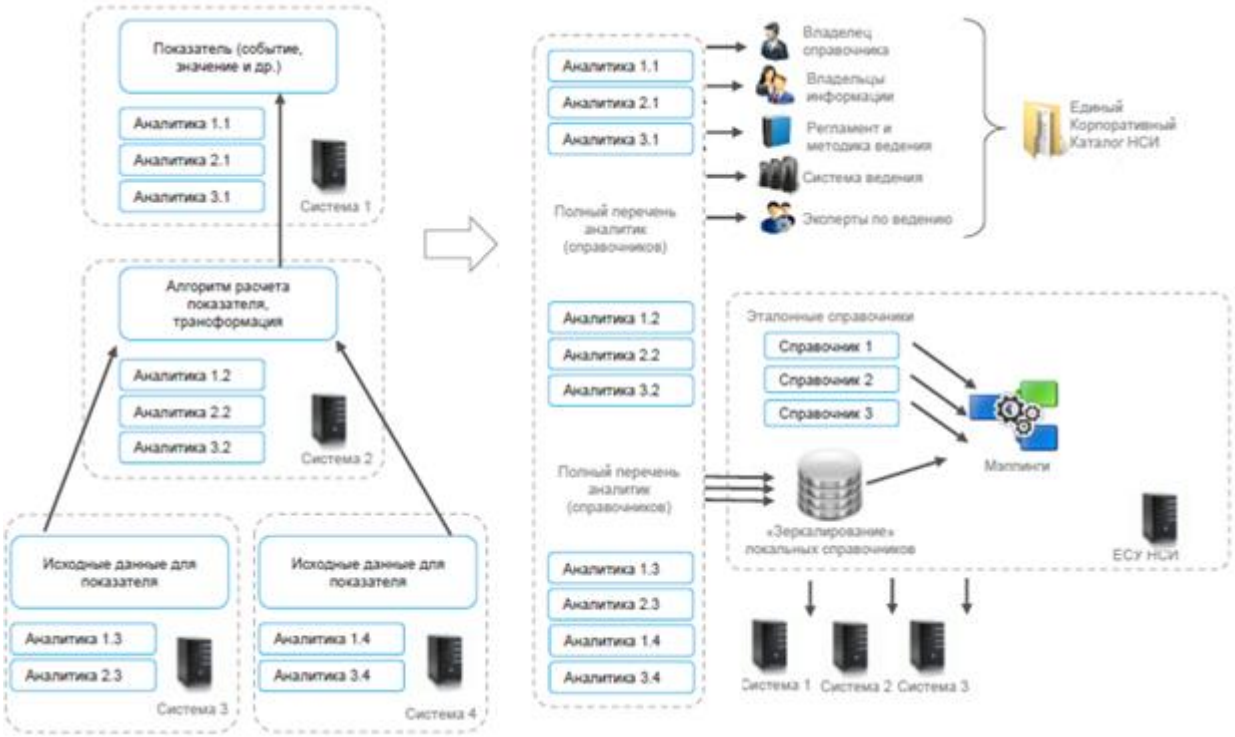

Рисунок 4.4. Процесс управления НСИ

4.2 Налаживание процесса управления знаниями в компании

Важнейшей и заключительной стадией процесса аналитики является этап накопления и сохранения знаний в компании. Лучшие практики компании, знания и опыт сотрудников должны быть сохранены и доступны для повторного использования. От того насколько грамотна налажена работа по управлению знаниями зависит эффективность предприятия в целом.

Очень важно выстраивать правильное сотрудничество в компании для продвижения идей и инноваций. Анализ «больших данных» позволяет делать открытия в данных, приносящие значительную пользу для компании. Однако эффективность данного анализа окажется невысока, если сотрудники не смогут получать информацию об уже готовых решениях и будут использовать ресурсы компании для проведения аналогичных исследований. В данном аспекте первоочередной задачей становится сбор экспертизы в компании и создание базы знаний, способствующей накоплению лучших практик и готовых решений.

Также важнейшим элементом данной работы является качественное предоставление данных. Сотрудники, участвующие в анализе, должны пользоваться единственным источником актуальных данных. В данном аспекте первоочередную роль играет единая система НСИ, о чем уже было сказано выше.

В то же время значительное влияние на процесс накопления знаний оказывает практика управления данными в компании. Для того, чтобы избежать ошибок, дубликатов, избыточности и несоответствия в данных необходима постоянная работа по поддержанию качества данных.

При правильном балансе самостоятельной аналитики и регулировании доступа к информации, компании способны стимулировать накоплении коллективных знаний в рамках всей организации, предоставлять готовые решения для различных команд аналитиков внутри компании, а также поощрять дискуссии и внедрение инноваций.

Таким образом, грамотный процесс накопления знаний может быть описан следующим образом:

1. Сбор знаний - сбор неформализованных знаний, проверка, добавление структуры
2. Обогащение знаний - проставление связей, формализация, добавление контекста
3. Запись в базу знаний
4. Анализ полученных знаний
    a. Оценка ценности знаний
    b. Разработка плана использования

5. Использование знаний
    a. Обеспечение доступности знаний
    b. Внедрение лучших практик
    c. Изменение бизнес-процессов

Улучшенный процесс аналитики в компании можно отобразить в виде схемы, представленной на рисунке 4.5. Как мы можем заметить процесс аналитики становится замкнутым, что позволяет накапливать наиболее ценную для компании информацию в виде знаний и лучших практик, записанных в хранилище данных. При наличии единой системы НСИ, проблемы на этапе загрузки и интеллектуального анализа данных также могут быть преодолены.

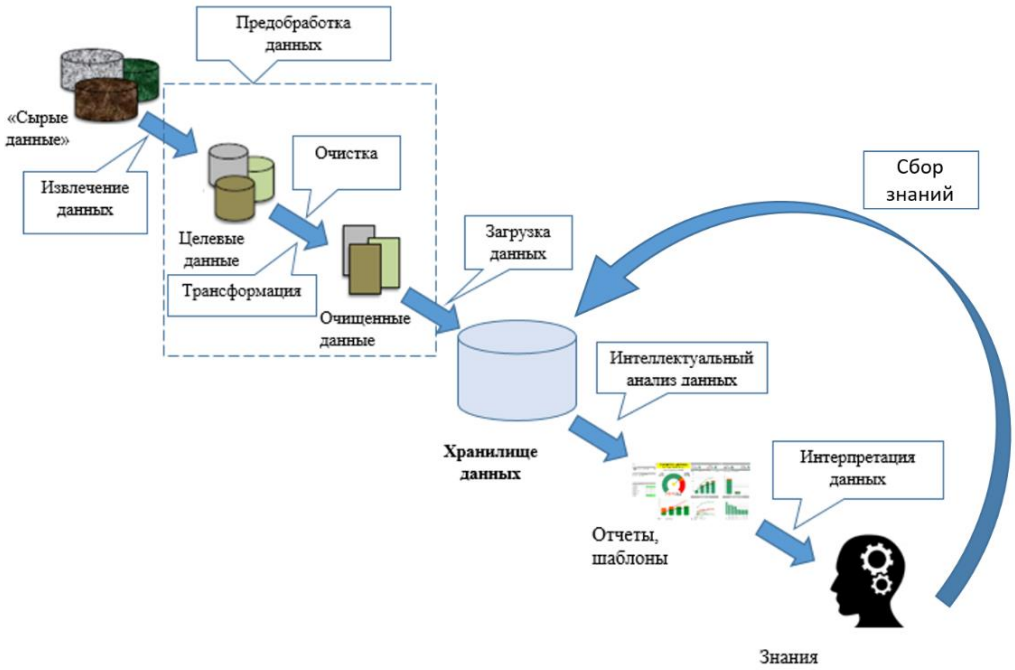

Рисунок 4.5. Улучшенный процесс аналитики в компании

Данные, которые приносят настоящую ценность бизнесу, как уже было определено в теоретической части данной работы и представляют собой Smart Data. Такие данные могут быть повторно использованы и при этом не потерять своей ценности, так как позволяют бизнес-пользователям и руководству предприятия выйти на более высокий уровень понимания бизнеса и открыть новые способы повышения его эффективности. В этой связи возникает необходимость в налаживании постоянного процесса гибкой аналитики, который заключается в более осознанном и разумном сборе данных и представляет собой реализацию концепции Smart Data.

Поддержка взаимодействия и сотрудничества, процесс накопления лучших практик и правила реагирования на инциденты формируют единые правила работы в компании, которые должны быть понятны всем сотрудникам.

Таким образом, для поддержания эффективности работы предприятия необходимо наладить постоянный процесс сбора и применения полученных знаний. Процесс накопления знаний не должен быть стихийным. Только в этом случае сотрудники компании смогут извлечь выгоду из имеющихся знаний, не исследуя заново то, что уже было открыто.

4.3 Разработка приложения, реализующего концепцию Smart Data

Одним из возможных решений возникающих проблем, связанных с обработкой и анализом «больших данных», является использование инструментов Business Intelligence (бизнес-анализа информации). Так, например, в данном контексте Smart Data выступают как данные, представленные таким образом, что их напрямую можно использовать для принятия управленческих решений или поддержания бизнес-процессов. Таким образом, визуализированная информация, графики и диаграммы, построенные на основе «больших данных», отвечают на запросы бизнес-пользователей и позволяют применять в практической деятельности результаты визуализации.

В настоящее время существует целый ряд компаний, предоставляющие свои системы бизнес-аналитики. Среди них можно выделить три продукта, которые традиционно входят в квадрант лидеров по версии Gartner [42]:

1. Microsoft Power BI
2. Tableu
3. QlikView

В компании, в которой была проведена практическая часть настоящей выпускной квалификационной работы, внедрена система QlikView, поэтому в дальнейшем мы также будем пользоваться инструментами данной системы. Исходя из формулировки поставленной проблемы, выбор системы QlikView является оптимальным и отлично подходит для решения задач данной выпускной квалификационной работы.

Среди достоинств QlikView можно выделить следующие:

- Возможности работы с данными из различных источников (что актуально при работе с «большими данными», а также при наличии большого числа систем источников)
- Большое количество учебных материалов, кейсов и ресурсов для самостоятельного изучения
- Обширное Qlik сообщество

Система бизнес-анализа QlikView способна решать множество проблем, возникающих в современном бизнес-анализе. Так, например, инструменты данной системы позволяют расширить возможности каждого работника организации, участвующего в совместной работе по поиску ценности в «больших данных». Система QlikView обеспечивает вовлечение в процесс анализа данных большого числа рядовых пользователей, глубоко не владеющих техническими аспектами работы с данными. В связи с этим, анализ данных перестает быть привилегией исключительно ИТ-специалистов.

При анализе данных значительная часть ценности возникает, когда компании объединяют данные из нескольких источников, обогащая таким образом имеющиеся данные и усиливая последующий анализ и понимание. Это относится и к «большим», и к традиционным данным.

Одной из самых актуальных проблем, с которой столкнулись ИТ-специалисты, является постоянная борьба за «поиск иголки в стоге сена». Системы бизнес-аналитики, в свою очередь делают процесс бизнес-анализа доступнее, так как все больше людей начинают экспериментировать со своими данными, что в конечном счете приводит к поиску скрытых зависимостей и извлечению дополнительной ценности. Чем больше людей участвует в анализе данных, тем больше генерируется идей по поводу данных, тем больше растет извлекаемая ценность и тем сильнее растет окупаемость вложений в «большие данные».

Подход QlikView состоит в том, чтобы расширять возможности всех бизнес-пользователей вне зависимости от их уровня владения техническими навыками. Таким образом, данная система предоставляет возможности простого использования «больших данных» абсолютно всеми группами пользователей.

Возможности QlikView это уникальная комбинация методов и средств:

1. Полный обзор всей информации компании – инструменты данной системы позволяют легко проводить интеграцию данных из различных источников, чтобы обеспечить наиболее полную картину бизнеса

2. Интерактивный опыт взаимодействия с информацией – пользователи могут выбирать и изменять варианты представления данных, в результате чего все визуализации обновляются немедленно
3. Множественные методы поддержки «больших данных»
    a. In-memory обработка запросов
    b. Сегментация и формирование цепочек
    c. Прямое обнаружение
    d. Разработка приложений по запросу

Данные возможности представлены на рисунке 4.6.

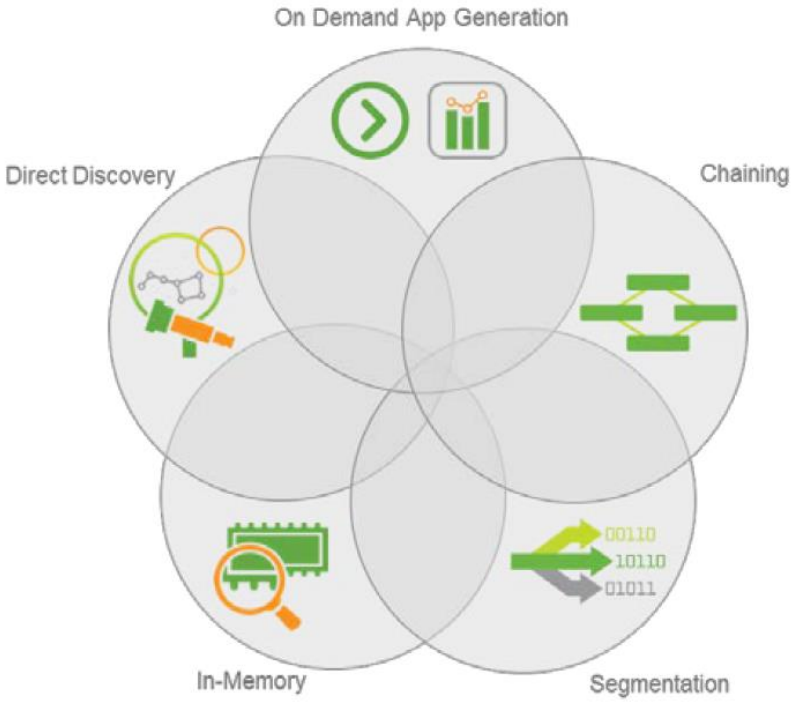

Рисунок 4.6. Qlik Big Data Methodologies

Ссылка на источник: https://www.qlik.com

Таким образом, QlikView позволяет расширить аналитику «больших данных» до границ всей компании, а также предоставляет широкие возможности анализа данных для нетехнических пользователей. Все это в совокупности позволяет рядовым пользователям без особых проблем проверять гипотезы и получать ценную информацию из «больших данных».

Рассмотрим пример практического применения модели Smart Data в бизнес-анализе. Система QlikView позволяет генерировать лог-файлы (потоковые данные) по каждому разработанному приложению (файлы формата .qvw).

Лог-файлы представляют собой файлы формата .txt и на рисунке 4.7 представлен пример структуры данного файла.

Exe Type	Exe Version	Server Started	Timestamp	Document
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 11:21:18	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 11:33:18	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 11:33:18	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 12:11:18	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 12:43:37	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 13:27:20	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 13:38:19	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 13:39:19	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 13:55:19	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 13:56:19	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 14:00:20	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 14:13:13	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 14:42:43	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 14:46:20	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 15:01:20	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 15:01:20	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 15:34:41	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 15:42:21	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 16:15:21	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 18:13:23	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 18:20:46	
RLS64	11.20.13603.0409.10	2017-08-29 10:37:09	2017-08-31 20:13:06	

Рисунок 4.7. Структура лог-файла

Как можно заметить, обычному бизнес-пользователю совершенно неочевидно, что обозначают эти данные, тем более неясно, как можно проводить аналитику по этим данным. Тем не менее, лог-файлы содержат в себе очень ценную информацию, которая может быть использована в дальнейшем при принятии управленческих решений.

Особенности потоковых данных заключаются в том, что они формируются непрерывно и отправляются небольшими объемами. Эти порции данных должны быть обработаны последовательно и инкрементально, после чего они уже пригодны для решения различных аналитических задач.

Задача бизнес-аналитика заключается в поиске ответов на следующие вопросы:

- Как представить данные в удобном, воспринимаемом виде?
- Как обеспечить непрерывную обработку поступающих потоковых данных?
- Как понятным способом донести основные идеи, возникающие по поводу данных?
- Как увеличить бизнес-ценность данных для пользователей?

- Как с помощью анализа лог-файлов приложения оценить его эффективность и удобство для пользователя?
- Как удовлетворить потребности пользователей наилучшим способом?

В то же время инструментарий разрабатываемого приложения должен предоставлять доступ для анализа данных широкому кругу пользователей, в частности пользователям из бизнеса, разбирающихся в специфике деятельности компании.

Так, разрабатываемое приложение должно удовлетворить следующие информационные потребности пользователей:

1. Потребность в более широком доступе к информации
2. Потребность в своевременной аналитике
3. Потребность в качественных данных для получения более точных результатов анализа

Данные должны быть предоставлены для аналитики в любое время, в любом месте, в рамках любого контекста или устройства.

Для решения всех поставленных задач было разработано приложение, предоставляющее пользователям доступ к продвинутой аналитике потоковых данных. Данное приложение реализует концепцию Smart Data, представляя потоковые данные в удобном и воспринимаемом виде. На рисунке 4.8 представлена модель данных разработанного приложения. С помощью инструментов QlikView потоковые данные были трансформированы и представлены в виде ассоциативной модели.

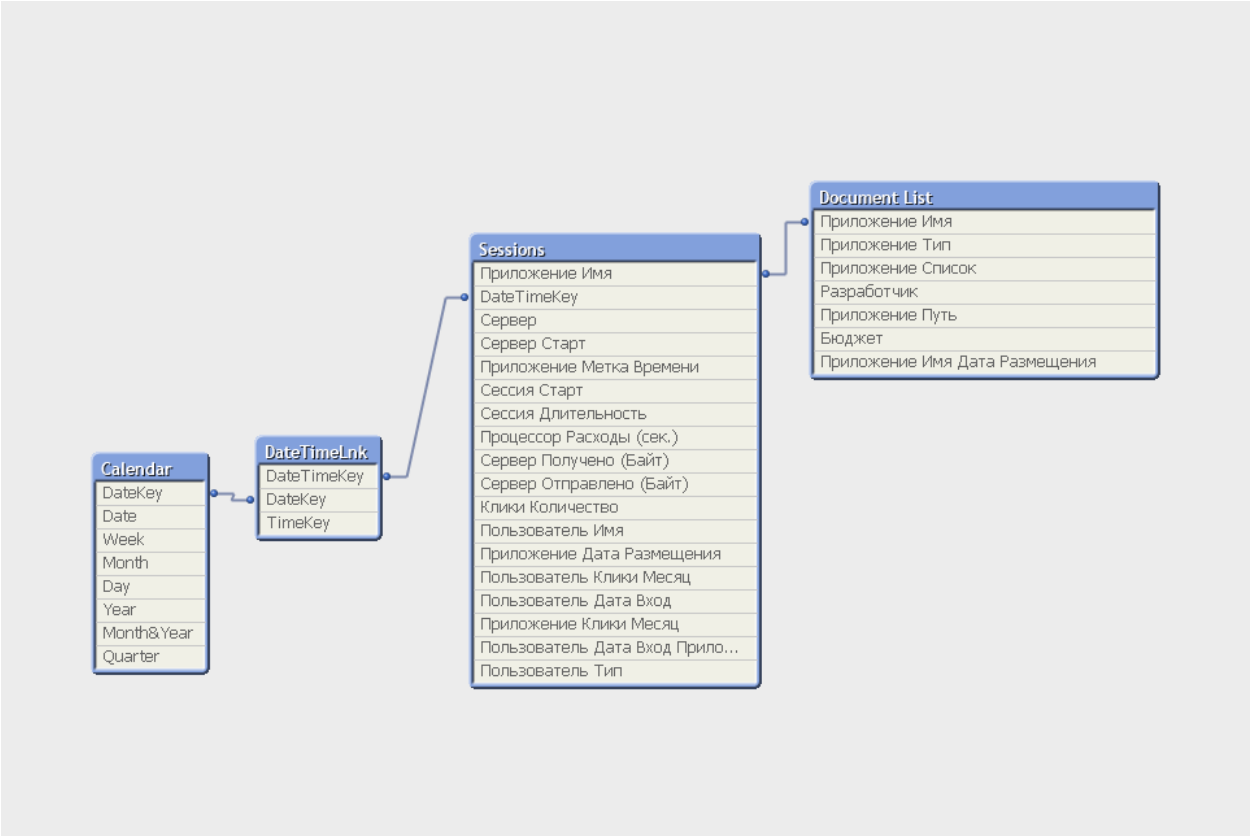

Рисунок 4.8. Модель данных приложения

Однако в таком виде потоковые данные всё ещё недостаточно понятны для обычных пользователей.

Вследствие этого на рисунке 4.9 представлена практическая реализация концепции Smart Data в виде разработанного приложения. В данном аспекте диаграммы и графики являются формой представления потоковых данных. В процессе интерпретации бизнес-аналитик способен вывести Smart Data, которые будут использованы в дальнейшем для принятия бизнес-решений и улучшения процесса аналитики.

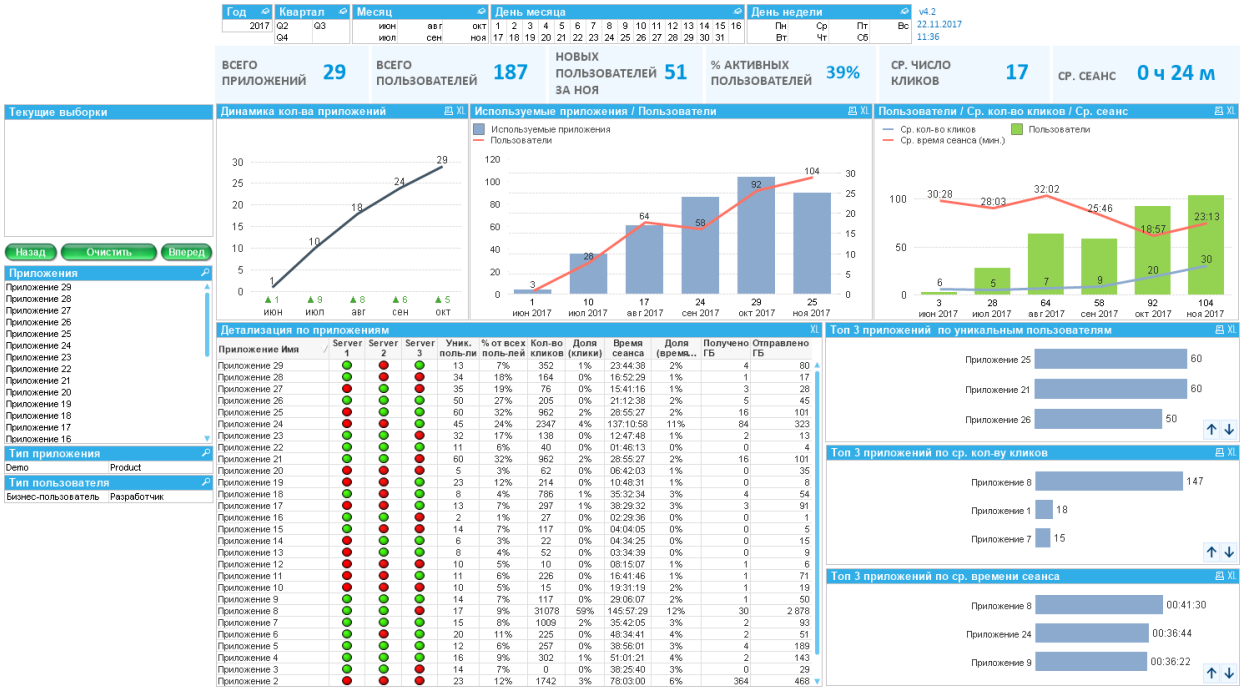

Рисунок 4.9. Приложение, реализующее концепцию Smart Data

С помощью инструментов QlikView бизнес-пользователь способен обработать и визуализировать данные лог-файлов, что является основой ценности, извлекаемой из поступающих данных.

Разработанное приложение позволяет бизнес-пользователям решить следующие задачи:

- Получить информацию об общем количестве уникальных пользователей целевого приложения
- Получить информацию о динамике активности уникальных пользователей
- Определить процент активных пользователей
- Определить среднее время сеанса и среднее число кликов по всем пользователям
- Получить возможность анализа данных без дополнительных знаний технических аспектов работы системы
- Оценить эффективность разрабатываемого целевого приложения
- Осуществлять на постоянной основе мониторинг удобства и эффективности работы пользователей при публикации новых версий целевого приложения
- Оценить удовлетворенность пользователей при публикации новых версий целевого приложения
- Определить какие данные представляют наибольшую ценность и должны быть отобраны в первую очередь в соответствии с концепцией Smart Data

Таким образом, выгоды от разработки данного приложения состоят в следующем:

1. Улучшение качества сервиса и предоставляемых услуг
2. Снижение риска потери клиента
3. Снижение риска провала проекта
4. Выявление информационных потребностей пользователей
5. Выявление удовлетворенности пользователей на основе анализа их поведения

В совокупности это позволит руководству предприятия проводить мониторинг и контроль эффективности разработки «Системы аналитической отчетности по цепочке добавленной стоимости» QlikView.

4.4 Оценка экономической эффективности

Разработанное приложение позволило в период с 01.01.2018 по 30.04.2018 снизить количество разрабатываемых релизов целевого приложения с 4-х до 3-х, а также привело к сокращению трудозатрат разработчиков и руководителя проекта для выпуска новой версии приложения на 15%.

Прежние трудозатраты на разработку 1-го релиза целевого приложения составляли 160 ч работы разработчика и 40 ч работы руководителя. Исходя из средней зарплаты данных специалистов, взятых с сайта HeadHunter [50], стоимость 1 ч работы для руководителя составляет 1250 руб., а для разработчика - 437.5 руб.

Таким образом, прямой экономический эффект в период с 01.01.2018 по 30.04.2018 от разработанного приложения составил:

$$\text{Экономический эффект (руб.)} = (40 + 40 \times 3 \times 0.15) \times 1250 + (160 + 160 \times 3 \times 0.15) \times 437.5 = 72500 + 101500 = 174000$$

Следует отметить, что помимо прямого экономического эффекта, также были получены и косвенные экономические выгоды от использования разработанного приложения:

- Снизились риски провала проекта внедрения целевого приложения
- Улучшилось качество принимаемых управленческих решений
- Сократилось время принятия управленческого решения
- Сократилось время разработки приложения
- Снизилось количество внедряемых версий целевого приложения
- Улучшилось качество и эффективность целевого приложения
- Повысилась удовлетворённость пользователей

Таким образом, разработанное приложение, реализующее концепцию Smart Data, является экономически целесообразным и приносит реальную бизнес-ценность компании.

4.5 Выводы

В практической части данной работы был проведён анализ компании, рассмотрена её сфера деятельности, исследован процесс аналитики и информационно-технологическая инфраструктура. На основании полученных данных были выявлены текущие проблемы компании и предложены рекомендации по их устранению.

Таким образом, в процессе проведения данной работы были получены следующие результаты:

- Предложены подходы по решению проблем аналитики в компании
- Предложены подходы по решению проблем управления данными в компании
- Предложено решение проблемы интеграции большого числа источников данных с помощью разработки единой системы НСИ
- Предложена методика формирования базы знаний компании
- Разработано приложение QlikView для анализа потоковых данных целевого приложения в соответствии с концепцией Smart Data

5 Заключение

В процессе проведения данного научного исследования были достигнуты все цели и задачи, поставленные во введении настоящей выпускной квалификационной работы. В рамках данной научной работы также были решены частные задачи на конкретном предприятии и предложены рекомендации для улучшения общего процесса аналитики.

В результате проведения настоящей выпускной квалификационной работы были получены следующие результаты:

- На основании анализа современных научных работ выявлены нерешённые проблемы в области «больших данных»
- На основании анализа современных научных работ обоснована актуальность, теоретическая значимость и прикладная ценность исследований по тематике «Smart Data»
- Предложена методика проведения литературного обзора посредством анализа естественного языка в отобранных научных работах методом интеллектуального анализа текста (text mining)
- Предложены подходы по решению проблем процесса аналитики в конкретной компании
- Предложено решение проблемы интеграции большого числа источников данных в виде разработки единой системы НСИ
- Предложена методика формирования базы знаний на примере конкретной компании
- Разработано приложение QlikView, реализующее концепцию Smart Data и удовлетворяющее требованиям бизнеса
- Предложены подходы по оценке эффективности разрабатываемых версий целевого приложения на основе анализа его потоковых данных

Таким образом, результаты, полученные в данной научной работе, могут без значительных изменений быть использованы в компаниях с аналогичным процессом аналитики и подобной информационно-технологической инфраструктурой.

Также следует отметить, что в процессе проведения литературного анализа были выявлены проблемы, которые получили недостаточную освещённость в современных научных исследованиях. Знание о существовании данных проблем могут быть в дальнейшем использованы исследователями для проведения последующих научных работ по тематике «больших данных».

6 Список использованной литературы

Статьи в журналах

1. Gantz J., Reinsel D. Extracting value from chaos //IDC iview. – 2011. – Т. 1142. – №. 2011. – С. 1-12.

2. Manyika J. et al. Big Data: The next frontier for innovation, competition, and productivity. – 2011.

3. Laney D. 3D data management: Controlling data volume, velocity and variety //META Group Research Note. – 2001. – Т. 6. – С. 70.

4. Boell S. K., Cecez-Kecmanovic D. A hermeneutic approach for conducting literature reviews and literature searches //Communications of the Association for Information Systems. – 2014. – Т. 34. – №. 1. – С. 257-286.

5. Wang Y. Business intelligence and analytics education: Hermeneutic literature review and future directions in is education //Browser Download This Paper. – 2015.

6. Chen M., Mao S., Liu Y. Big Data: A survey //Mobile Networks and Applications. – 2014. – Т. 19. – №. 2. – С. 171-209.

7. Chen H., Chiang R. H. L., Storey V. C. Business intelligence and analytics: From Big Data to big impact //MIS quarterly. – 2012. – Т. 36. – №. 4. – С. 1165-1188.

8. Wu X. et al. Data mining with Big Data //ieee transactions on knowledge and data engineering. – 2014. – Т. 26. – №. 1. – С. 97-107.

9. Chen C. L. P., Zhang C. Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data //Information Sciences. – 2014. – Т. 275. – С. 314-347.

10. Xu X. From cloud computing to cloud manufacturing //Robotics and computer-integrated manufacturing. – 2012. – Т. 28. – №. 1. – С. 75-86.

11. Zanella A. et al. Internet of things for smart cities //IEEE Internet of Things journal. – 2014. – Т. 1. – №. 1. – С. 22-32.

12. Cambria E. et al. New avenues in opinion mining and sentiment analysis //IEEE Intelligent Systems. – 2013. – Т. 28. – №. 2. – С. 15-21.

13. Bengio Y., Courville A., Vincent P. Representation learning: A review and new perspectives //IEEE transactions on pattern analysis and machine intelligence. – 2013. – Т. 35. – №. 8. – С. 1798-1828.

14. Hashem I. A. T. et al. The rise of “Big Data” on cloud computing: Review and open research issues //Information Systems. – 2015. – Т. 47. – С. 98-115.

15. Bakshy E. et al. The role of social networks in information diffusion //Proceedings of the 21st international conference on World Wide Web. – ACM, 2012. – С. 519-528.

16. Zissis D., Lekkas D. Addressing cloud computing security issues //Future Generation computer systems. – 2012. – Т. 28. – №. 3. – С. 583-592.

17. Venkatesh V., Thong J. Y. L., Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. – 2012.

18. Rosseel Y. lavaan: an R package for structural equation modeling and more Version 0.4-9 (BETA) //Retrieved from. – 2011.

19. Uijlings J. R. R. et al. Selective search for object recognition //International journal of computer vision. – 2013. – Т. 104. – №. 2. – С. 154-171.

20. Akay B., Karaboga D. A modified artificial bee colony algorithm for real-parameter optimization //Information Sciences. – 2012. – Т. 192. – С. 120-142.

21. Dinh H. T. et al. A survey of mobile cloud computing: architecture, applications, and approaches //Wireless communications and mobile computing. – 2013. – Т. 13. – №. 18. – С. 1587-1611.

22. Huang G. B. et al. Extreme learning machine for regression and multiclass classification //IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). – 2012. – Т. 42. – №. 2. – С. 513-529.

23. Blei D. M. Probabilistic topic models //Communications of the ACM. – 2012. – Т. 55. – №. 4. – С. 77-84.

24. Bobadilla J. et al. Recommender systems survey //Knowledge-based systems. – 2013. – Т. 46. – С. 109-132.

25. Beloglazov A., Abawajy J., Buyya R. Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing //Future generation computer systems. – 2012. – Т. 28. – №. 5. – С. 755-768.

26. De Brito M. A. G. et al. Evaluation of the main MPPT techniques for photovoltaic applications //IEEE transactions on industrial electronics. – 2013. – Т. 60. – №. 3. – С. 1156-1167.

27. Boccardi F. et al. Five disruptive technology directions for 5G //IEEE Communications Magazine. – 2014. – Т. 52. – №. 2. – С. 74-80.

28. Gubbi J. et al. Internet of Things (IoT): A vision, architectural elements, and future directions //Future generation computer systems. – 2013. – Т. 29. – №. 7. – С. 1645-1660.

29. Dollar P. et al. Pedestrian detection: An evaluation of the state of the art //IEEE transactions on pattern analysis and machine intelligence. – 2012. – Т. 34. – №. 4. – С. 743-761.

30. FROM BIG DATA TO SMART DATA: Using data to drive personalized rand experiences Rob Salkowitz, January 22, 2014

31. Королев О. Л., Апатова Н. В., Круликовский А. П. «Большие данные» как фактор изменения процессов принятия решений в экономике //Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Экономические науки. – 2017. – Т. 10. – №. 4.

32. Гродзенский С. Я., Калачева Е. А. Большие данные: история, перспективы, потенциал //Стандарты и качество. – 2017. – №. 8. – С. 64-67.

33. Sizov I. BIG DATA–БОЛЬШИЕ ДАННЫЕ В БИЗНЕСЕ //Экономика. Бизнес. Информатика. – 2017. – Т. 2. – №. 3.

34. Кравченко В. О., Крюкова А. А. «Большие данные»-практические аспекты и особенности //Academy. – 2016. – №. 6. – С. 65-67.

35. Зоткин А. С., Ворожцов А. С. БОЛЬШИЕ ДАННЫЕ: СОВРЕМЕННЫЕ ТЕХНОЛОГИИ ОБРАБОТКИ ИНФОРМАЦИИ //И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ. – 2016.

36. Шлюйкова Д. П. Большие данные: современные подходы к хранению и обработке //Наука, техника и образование. – 2016. – №. 1. – С. 75. – С. 81.

37. Черникова Е. И. Интернет вещей и технология big data (большие данные) //Ученые записки ИСГЗ. – 2017. – №. 1. – С. 581-584.

38. Назаренко Ю. Л. ОБЗОР ТЕХНОЛОГИИ «БОЛЬШИЕ ДАННЫЕ»(BIG DATA) И ПРОГРАММНО-АППАРАТНЫХ СРЕДСТВ, ПРИМЕНЯЕМЫХ ДЛЯ ИХ АНАЛИЗА И ОБРАБОТКИ //European Science. – 2017. – №. 9. – С. 25-30.

39. Ковалевский А. Е., Ефремов Е. А. БОЛЬШИЕ ДАННЫЕ //Новая наука: Стратегии и векторы развития. – 2016. – №. 6-1. – С. 27-28.

Интернет-ресурсы и электронные базы данных

40. Information Age [Электронный ресурс]: URL: http://www.information-age.com// (Дата обращения: 02.10.2016)

41. Wired [Электронный ресурс]: URL: http://www.wired.com//(Дата обращения: 18.08.2017)

42. Gartner [Электронный ресурс]: URL: http://www.gartner.com//(Дата обращения: 02.10.2016)

43. RusBase[Электронный ресурс]: URL: http://rb.ru//(Дата обращения: 11.11.2016)

44. Big Data Bussiness Summit [Электронный ресурс]: URL: http://bigdatasummit.ru/bd2015//(Дата обращения: 11.11.2016)

45. Dataconomy [Электронный ресурс]: URL: http://dataconomy.com//(Дата обращения: 17.08.2017)

46. Forbes [Электронный ресурс]: URL: https://www.forbes.com//(Дата обращения: 18.08.2017)

47. Blue-Granite [Электронный ресурс]: URL: https://www.blue-granite.com //(Дата обращения: 17.04.2018)

48. Globenewswire [Электронный ресурс]: URL: https://globenewswire.com //(Дата обращения: 17.04.2018)

49. Habrahabr [Электронный ресурс]: URL: https://habrahabr.ru //(Дата обращения: 17.04.2018)

50. HeadHunter [Электронный ресурс]: URL: https://hh.ru //(Дата обращения: 02.05.2018)

Приложение 1

Таблица «Список сокращений»

Сокращение	**Расшифровка**
БД	База данных
СУБД	Система управления базами данных
ИС	Информационная система
ИТ	Информационные технологии
ПО	Программное обеспечение
ХД	Хранилище данных
НСИ	Нормативно-справочная информация
SQL	Script Query Language
NoSQL	Not only SQL
IoT	Internet of things
CDO	Chief Data Officer
API	Application programming interface

Аннотация

***Ключевые слова**: big data, smart data, business intelligence, большие данные, интеллектуальные данные, бизнес-анализ информации*

В настоящей выпускной квалификационной работе были рассмотрены основные направления развития научных исследований по теме «большие данные», выделены современные противоречия теории и практики, а также обоснована актуальность исследований по данной тематике. На основании современных тенденций развития «больших данных» в рамках настоящей научной работы была рассмотрена концепция Smart Data. Цель данной работы заключалась в разработке рекомендаций для использования концепции Smart Data в практической деятельности компаний.

В результате проведения настоящих исследований были предложены рекомендации по решению проблем, связанных с процессом аналитики, менеджментом данных, а также процессом управления знаниями. Предложенные рекомендации являются практически значимыми и могут быть применены без значительных изменений в аналогичных компаниях.